THE UNIVERSITY
of EDINBURGH

Bioinformatic analysis of genome-scale data reveals insights into host-pathogen interactions in farm animals

Mick Watson

PhD by Research Publication

The Roslin Institute and Royal (Dick) School of Veterinary Studies

University of Edinburgh

2015

Declaration

I declare that this thesis and the work presented therein is my own work, except where explicitly stated. This work has not been submitted for any other degree of professional qualification

Mick Watson, 2015

Abstract

This thesis documents the contribution of my bioinformatics research activities, including novel software development, to a range of research projects aimed at investigating the interactions between bacterial and viral pathogens and their hosts. The focus is largely on farm animal species and their pathogens, although some of the research has a wider scientific impact.

RNA interference (RNAi) refers to a variety of related regulatory pathways present in animals, plants and insects. The major pathways are microRNAs (miRNAs), small-interfering RNAs (siRNAs) and PIWI-interacting RNAs (piRNAs). Marek's disease virus is an important pathogen of poultry, causing T-cell lymphoma. We identified the presence and expression patterns of several MDV-encoded microRNAs, including the identification of 5 novel microRNAs. We also showed that not only do virus-encoded microRNAs dominate the mirNome within chicken cells, but also that specific host-microRNAs are down-regulated. We also identify novel virus-encoded microRNAs in other *Herpesviridae* and provide the first evidence of miRNA evolution by duplication in viruses. In related work, we present a novel microRNA generated by the canonical miRNA biogenesis pathway in Avian Leukosis Virus, another avian oncogenic virus, and publish data showing the expression pattern of known chicken microRNAs across a range of important avian cells. Two of the other RNAi pathways (siRNA and piRNA) form an important part of the antiviral response in arthropods. We have published work demonstrating an siRNA antiviral response to bluetongue virus and Schmallenberg virus in cells from the *Culicoides* midge, an important insect vector, as well as work demonstrating the importance of the piRNA pathway in the antiviral response to Semliki forest virus (SFV). Further work on flaviviruses in ticks demonstrates the active suppression of the siRNA response by Langat Virus, as well as a key difference between the siRNA responses in Mosquitos compared to ticks.

Salmonella is one of the most important zoonoses, with an estimated 1.4 million cases of human salmonellosis per annum in the USA alone. Salmonella infections of farm animals are an important route into the human food chain. This thesis presents work on the comparative structure and function of 13 fimbrial operons within *Salmonella enterica* serovar Enteritidis as well as a genomic comparison of that serovar with *Salmonella enterica* serovar Gallinarum, a chicken-specific serovar. We characterised the global expression profile of *Salmonella enterica* serovar Typhimurium during colonization of the

chicken intestine, and we have published the genomes of four strains of *Salmonella eneterica* serovars of well-defined virulence in food-producing animals. Our work in this area led to us publishing an important and comprehensive review of the automatic annotation of bacterial genomes.

Finally, I present work on novel software development. ProGenExpress, a software tool that allows the easy and accurate integration and visualisation of quantitative data with the genome annotation of bacteria; Meta4 is a web application that allows data sharing of bacterial genome annotations from metagenomes; CORNA, a software tool that allows scientists to link together microRNA targets, gene expression and functional annotation; viRome, a software tool for the analysis of siRNA and piRNA responses in virus-infection studies; DetectiV, a software tool for the analysis of pathogen-detection microarray data; and poRe, a software tool that enables users to organise and analyse nanopore sequencing data

Acknowledgements

I would like to thank The Roslin Institute and the University of Edinburgh for giving me the opportunity to gain a PhD by research publication.

This thesis covers my career since 2002, and in that time I have been influenced by quite a few colleagues. At the Institute for Animal Health (now The Pirbright Institute), I would like to thank Geoff Oldham and Paul Pierre Pastoret for giving me the opportunity to begin my own research group. Whilst at IAH, Nat Bumstead, Martin Shirley, David Paton and Fiona Tomley all provided advice and stewardship. In particular I would like to thank Fiona, who provided the perfect mix of support, encouragement and scientific challenge.

I owe a great deal of debt to The Roslin Institute for giving me the opportunity to continue my research career. Both Alan Archibald and David Hume have provided advice, guidance and yet more challenges, and have allowed me to build my research career further.

I have enjoyed a long term collaboration with Prof Venu Nair which has been amazingly fruitful in terms of research outputs.

I would also like to thank the funders who have funded my work, including BBSRC, TSB, the EU, and Defra.

Contents

1. Introduction

In this thesis I present my work from over 12 years spent in UK academia, working as a bioinformatics researcher in the field of farm animal health and food security. Naturally, as a bioinformatician, my work has focused on the analysis of data from a variety of genomics and post-genomics technologies. The pace-of-change in genomics has been incredible, and in my career, spanning 16 years in industry and academia, I have witnessed technologies rise and subsequently die as other, better technologies come along. Data types and sizes have also changed an incredible amount during my career, and problems that would have been impossible just a few years ago now seem routine. Of course, this is because advances in computing have kept pace with advances in genomics: disk storage is much cheaper, the internet is faster, as are computer processors. There have been advances in parallel computing and the rise of cloud computing. All of these have helped us cope with the pace of change in genomics and bioinformatics.

Bioinformatics, by its very nature, is a collaborative science, and it is my opinion that the very best science is carried out when multidisciplinary teams come together, with complementary expertise, to tackle a problem. I present in this thesis a range of collaborative research papers and projects. In some of them, I have been the major driver of that research, as either first or last author. In others, I was simply part of the team that produced the research. A list of the papers included in the thesis, and my contribution to each one, is given in appendix I.

As a bioinformatician, one of the key outputs of my research has been open-source software, written and released for others to use, and I describe some of these software tools in section 4. The main driver behind the development of these tools has always been that there is an unmet need in the community, and often I became familiar with that unmet need through collaboration with researchers. Many of the collaborative research papers I present in this thesis are related to the software tools in section 4 – this is because the methods used in the software were developed through collaboration with researchers who needed those methods. The outputs of my research activities are often both new knowledge and new software, both of which have been published. Some of the software tools have been published as a short “Application Note” in the journal Bioinformatics, and these are limited to two pages in length. However, their short nature should not be taken as an indication that they are of lesser significance than longer papers. Application Notes

are reviewed as thoroughly as any other paper, and much of the hard work is "hidden" – not only must the software be written, but also documentation and tutorials which appear online.

I began my career in industry, in 1997 which was prior to the publication of the human genome. Bioinformatics was in its infancy – tools such as GCG, ClustalW and BLAST were in common use, but genome-scale experiments and data sets were rare. I began my academic career at the Institute for Animal Health in 2002, after the human genome was published, but prior to the publication of the chicken [1], cow [2], pig [3] and sheep genomes[4]. Much of my work focused on the genomics and transcriptomics of important farm animal pathogens, such as *Salmonella* and Marek's disease. At that point, genome sequencing was still only carried out in large sequencing centres. Much of the effort in bioinformatics outside of these centres was focused on analysis of post-genomics data, such as microarray data. The Bioconductor project [5] really launched R (http://www.r-project.org) as a platform for bioinformatics research, a platform which I have used extensively throughout my research career. As times and technologies changed, next-generation sequencing began to come to the fore, first 454 and Illumina/Solexa, followed by Ion Torrent, Pacific Biosciences and Oxford Nanopore. The pace of change in sequencing is frightening, and the impact on the bioinformatics community has been huge. Now more than ever, bioinformatics is key to genomics research, and the new technologies have driven a huge wave of novel bioinformatics research.

Any body of research needs to be set within the wider context, and the context of this thesis is food security. For mankind, feeding our species is already a problem, with some estimating that over 1 billion people worldwide are hungry (http://www.who.int/mediacentre/factsheets/fs311/en/). In the next few decades, consumption of meat is set to double as the human population increases. We therefore need to invest heavily into research that enables us to feed more people for less. This focus sits at the heart of both research institutes I have worked at, The Institute for Animal Health (now The Pirbright Institute) and The Roslin Institute, University of Edinburgh. The body of work described below, and in the attached papers, contributes to the knowledge of how farm animals and their pathogens function, and how they interact during both health and disease.

2. RNA interference in viral disease

RNA interference (RNAi) is a general term referring to a range of biological processes in which gene expression is inhibited by RNA molecules, typically via the targeting and destruction of mRNA molecules. RNAi was first discovered in *C elegans* [6] after the observation that dsRNA was a more potent suppressor of gene activity than either sense strand individually. Whilst RNA-mediated gene-silencing had been used extensively beforehand (e.g. [7]), it had been thought that this operated via complementary base-pairing of mRNA, thus blocking translation. Whilst initially the mechanism of RNAi was unknown, studies of the genes required for RNAi in *C elegans, Drosophila,* plants and funghi [8, 9] revealed a common underlying mechanism. Small RNAs were identified as an important agent in RNAi, at first in plants and then in other organisms [10, 11]. The discovery and characterisation of microRNAs [12, 13] linked the emerging world of RNAi with a natural RNA-mediated gene regulatory mechanism, leading to the common model we understand today whereby short RNAs act as guides for gene silencing via the RNA-induced silencing complex (RISC) [14].

There are a range of pathways that contribute to RNAi, but only three are relevant to this thesis – small interfering RNAs (siRNA), microRNAs (miRNA) and Piwi-interacting RNAs (piRNA). My work has involved the study of these pathways in the context of viral infections of insects and farm animals, and I will describe each below.

2.1 Characterisation of microRNAs in a range of avian viruses

Marek's disease (MD) is induced by Marek's disease virus type 1 (MDV-1) and causes the rapid and aggressive onset of T-cell lymphoma of poultry. MDV-1 is a highly contagious alphaherpesvirus, and MD is a major source of economic loss in the poultry industry worldwide. The major oncogenic protein encoded by the MDV-1 genome is considered to be Meq [15]; however, the genome also contains genes for several microRNAs. Herpesviruses have exploited microRNAs most successfully, with the vast majority of virus-encoded microRNAs in miRBase [16] coming from this family. There is also a large body of literature on microRNAs and how they contribute to oncogenesis (reviewed in [17]).

In [18], we investigated the expression of microRNAs in MSB-1, an MDV-transformed CD4+ T-Cell line derived from an MDV-1 induced lymphoma [19]. In this paper, a cDNA library was constructed from small RNAs extracted from MSB-1 cells, identifying a total of 5099

high quality sequence reads. Of these, 1641 mapped to known or novel chicken microRNAs, 2562 to the MDV-1 Md5 strain genome and 518 to the MDV-2 genome (MSB-1 is co-infected with both MDV-1 and MDV-2). The reads mapping to MDV-1 included 8 existing microRNA genes, but also identified 4 novel microRNA genes. A further novel microRNA was discovered using Northern blotting. Cloning frequency and Northern blotting were further used to characterise the expression of all 13 MDV-1 encoded microRNAs. The major findings of this work were that (i) more than 50% of the mirNome of MSB-1 cells is derived from MDV-1, with a further 10% derived from MDV-2; and (ii) the discovery of 5 novel MDV-1 microRNA genes. The high expression of certain microRNAs is an indication that they may be involved in regulating viral and host genes involved in infected/transformed cell lines. Whilst this study was an important advance in the field, the results must be considered in context. Findings in cell lines do not always translate directly in real biological systems – of note is that the expression of microRNAs we measured in this study did not always correlate well with a similar study in chick embryonic fibroblasts (CEF) [20]. Further issues include the use of cloning frequencies as a proxy for gene expression, and the co-infection of MSB-1 with both MDV-1 and MDV-2. In 2007/2008, when the study was carried out, next-generation sequencing was in its infancy. The study would clearly have benefitted from such an approach; indeed, Burnside and Morgan [20] used 454 pyrosequencing. Despite these problems, the study was the first of its kind using an avian lymphoma, and enabled the identification of known and novel microRNAs encoded by Marek's disease virus.

We carried out a follow up study in [21] using microarrays to study global changes in microRNA expression in Marek's disease virus-transformed cell lines. Here we measured the expression of host and viral microRNAs using a customised microarray. 7 MDV-1 transformed cell lines were used (including MSB-1). Reticuloendotheliosis virus T (REV-T) and avian leucosis virus (ALV) transformed cell lines were included as examples of MDV-negative virus transformed cell lines. Normal splenoctyes and CD4+ T-cells were used as controls. The results not only showed massive up-regulation of MDV encoded microRNAs (as would be expected; the control samples should not contain any MDV!), but also down-regulation of specific host microRNAs. By using uninfected splenoctyes, we were able to compare MDV-positive transformed cell lines with an MDV-negative transformed cell line, and this identified host microRNAs specifically down-regulated by MDV. Key amongst these was gga-miR-155, orthologues of which are known to be involved in oncogenesis in humans

and a regulator of the T-cell response [22]. Parallel work by Prof. Nair's group showed that MDV-miR-M4 is a functional orthologue of miR-155 and shares the same seed sequence [23]. Together, these studies raise the intriguing possibility that MDV specifically down-regulated host miR-155 whilst expressing large amounts of miR-M4, despite these two sharing a similar function. This study was the first of its kind to use microarrays to study global microRNA expression in Marek's disease transformed cell lines. We showed the relative expression of MDV encoded microRNAs, as well as the MDV specific down-regulation of specific microRNAs, including miR-155 a known regulator of oncogenesis and the immune response.

We went on to identify microRNAs in two other herpesviruses. In [24], we studied Herpesvirus of Turkeys (HVT), a virus from the same genus (*Mardivirus*) as MDV. HVT is widely used as a live vaccine against MDV. Using a similar approach to the one above, [18], we identified 11 novel microRNAs encoded within the HVT genome, 10 of which clustered together within a 2.1Kb window. The close proximity to one another, and certain groups having high homology to one another, provided evidence of evolution by gene duplication. This was a small study, but by identifying 11 novel HVT microRNAs we enabled further studies of their expression and function. The HVT microRNAs are also the first in viruses to be show evidence of evolution by duplication. We finally joined the next-generation sequencing revolution with a study in duck enteritis virus (DEV) [25]. In this study we used Illumina next-generation sequencing and microarrays to discover and characterise the expression of DEV microRNAs. We defined 24 genomic loci with high likelihood of being a novel microRNA gene. Microarray data comparing expression of 72 custom probes designed to putative microRNAs identified from the NGS data using DEV infected CEF labelled with Cy5 and uninfected CEF labelled as Cy3. By comparing to uninfected CEF, we are using background noise/signal/hybridisation as a control. Thus only microRNAs expressed at a greater level ($p<=0.01$) than background were considered to be "real" microRNAs. The microarray data also provided data on expression of host microRNAs, with 26 significantly up-regulated and 19 significantly down-regulated. By identifying 24 novel DEV-encoded microRNAs, we enabled further studies of microRNA function in a virus which causes a highly contagious and lethal disease in waterfowl.

In [26], we further used illumina next-generation sequencing technology to measure the expression of avian microRNAs in a range of haemopoietic cells. The purpose of this study

was to provide a background reference dataset of microRNA expression in different avian cell types. Included were 6 avian cell populations: BP25, a chickembryonic stemcell (cESC) line; Bu1B, naïve embryonic B lymphocytes; StimB, CD40L-induced B-cells; DT40, an avian-leukosis virus (ALV) transformed B-cell line; HD11, a chicken macrophage cell line; and IAH30, a turkey macrophage cell line. In addition to providing expression profiles for the individual cells/cell lines, comparison across the data set identified distinct clusters of microRNAs whose expression seemed to be limited to a single cell type/line. By comparing stimulated to un-stimulated B-cells, we were able to provide evidence of increased expression of a range of microRNAs, some of which had been previously indicated in cell proliferation. Again, gga-miR-155 played a prominent role, showing the largest fold-increase in expression upon stimulation. The data we published and released as part of this study will be useful to those researchers using the same or similar cell types/lines, and represent one of the first reference datasets for microRNA expression in chicken and turkey haemopoietic cells.

The data from IAH30 demanded further attention [27]. IAH30 is a turkey macrophage cell line transformed by avian leucosis virus (ALV) subgroup J. ALV is a retrovirus with a single stranded RNA genome (compared to the double-stranded DNA genomes of the Herpesviruses discussed above), and is known to induce a range of different types of cancer in poultry [28]. Of particular interest are a number of regulatory elements found to be encoded within the genomes of retroviruses, one example of which is the XSR region which has been identified in both Rous sarcoma virus (RSV) [29] and ALV [30]. The function of the XSR element is unknown, though it has been speculated that it acts as a transcriptional enhancer [31]. During our work for [26], we noticed a huge number of small RNA reads mapping to the XSR element of the ALV genome (accession Z46390.1). Indeed, of 1.5 million total reads in the library, just under 360000 mapped to a single locus in the XSR element (335 reads mapping to a region a short distance downstream). Only 580 reads mapped elsewhere on the ALV genome. The huge number of reads, and the characteristic expression pattern, led us to hypothesize that these reads came from a novel microRNA. The surrounding genomic sequence is also predicted to form a characteristic hairpin structure. Using Northern blotting, we compared IAH30 cell lines with those of REV-T transformed turkey cell line AVOL-1 and uninfected turkey spleen cells. The putative ALV XSR microRNA was shown to be expressed in IAH30 but in neither of the controls. As IAH30 is derived from macrophages, we also examined ALV XSR microRNA expression in

macrophage cells. Expression was found by TaqMan in two other cell lines infected with ALV type J viruses, and in a chicken blastoderm cell line infected with ALV. Expression was not detected in cells infected with ALV with the XSR region deleted or from macrophage cells infected with REV-T. Taken together, these provide significant evidence that the putative microRNA is derived specifically from the XSR element of ALV. By studying the genomes of ALV and RSV, we were able to align the XSR element from 64 viral isolates. Within those alignments, the putative 5' microRNA was highly conserved (identical in 46; 1 SNP in 16; 2 SNPs in 2). Transfection and knockdown studies confirm that expression of the ALV XSR microRNA is driven by the Pol II promoter, and the microRNA is processed by Drosha and Dicer. In further experiments we demonstrated that the ALV XSR microRNA is capable of significantly down-regulating the expression of RNA containing artificial target sites of perfect complementarity to the microRNA sequence.

Our study was one of the first to demonstrate microRNA expression from a single-stranded RNA virus, and the first to demonstrate this in ALV. It is also the first to provide evidence that a single stranded RNA virus can use the canonical microRNA processing pathway. The high level of ALV XSR microRNA expression in the IAH30 cell line indicates an important role in ALV pathogenesis.

2.2 siRNA and piRNA in insect vectors

Both small interfering RNAs (siRNA) and Piwi-interacting RNAs (piRNA) are thought to determine the specificity of RNA silencing by recognising complementary RNA or DNA targets. However, they are the result of distinct pathways: siRNAs are dependent on Dicer nucleases, whereas piRNAs are Dicer-independent [32-34]. Both siRNAs and piRNAs have been shown to be induced by virus infection in a wide range of hosts, and are thought to modulate a range of virus-host interactions. Virus-derived siRNA molecules act as guides for RNAi-based antiviral immunity in plants, invertebrates and funghi (reviewed in [32]). In animals, the primary role of piRNAs is thought to be the silencing of transposable elements in the germ-line. piRNAs associate with PIWI proteins to form piRNA-induced silencing complex (piRISC), that recognises and silences complementary targets (reviewed in [33]). However, discovery of virus-derived piRNAs in *Drosophila melanogaster* led to speculation that piRNAs may also play a role in antiviral immunity [35].

In collaboration with Alain Kohl's group, we have used next-generation sequencing to investigate the patterns of virus-derived siRNA and piRNA expression in mosquito, midge and tick cells, and I will describe these studies below.

The term "arbovirus" is used to refer to viruses that are transmitted by arthropod vectors, and biting arthropods (including mosquitos, ticks and midges) are an important vector of animal and human diseases. These include important human pathogens such as Dengue virus (DENV) and West Nile virus (WNV); and important animal pathogens, such as bluetongue virus (BTV) and Schmallenberg virus (SBV). Much of the research into the antiviral RNAi response to arboviruses has been carried out in mosquitos [36, 37], and our study was the first to investigate the siRNA response to arboiruses in a midge, specifically *Culicoides sonorensis* [38]. We investigated the siRNA response in *C. sonorensis*-derived KC cells and assessed whether the antiviral response targets arboviruses using both Bluetongue virus (BTV; which has a double-stranded RNA genome) and Schmallenberg virus (SBV; which has a single-stranded RNA genome). The study is particularly important as BTV and SBV are important animal pathogens: BTV is a notifiable disease within the EU, affecting all ruminants and causing acute disease. SBV is a recently emerged disease that can cause congenital malformations in newborns [39]. After infection with BTV or SBV, the RNAi response was measured using Illumina next-generation sequencing. The study showed that *Culicoides* cells can and do mount an RNAi response to both the dsRNA virus BTV and the ssRNA virus SBV, with 21nt RNA species being the dominant form of viRNA. Interestingly, there were also peaks at 25-29nt, suggestive of a piRNA response, though this was not investigated further. The induction of a siRNA antiviral response is a strong indicator that an exogenous RNAi pathway is present in *Culicoides* species. This work was the first to characterise the antiviral siRNA response in midges, and represents an important advance in the field.

As stated above, much of the arbovirus RNAi antiviral response research has focused on mosquitos. Ticks are another important vector of arboviruses, and we have characterised the role of key RNAi proteins in the response to Langat virus (LGTV) in an *Ixodes scapularis*-derived cell line [40]. In *Drosophila*, viRNAs integrate with the Ago-2 protein, and guide the degradation of target RNAs using sequence complementarity [41]. Putative Ago proteins have been identified in the *I. scapularis* genome, but their role in the antiviral response had not been characterised [42]. In our study, we used RNAi to knock down transcription of

specific Ago proteins in IDE8 cells, an *I scapularis* cell line, and characterised their effect on the LGTV replicon. By characterising the antiviral RNAi response, we were also able to show that tick siRNA are 22nt in length, compared to 21nt in *Drosophila* and *Culicoides*. Using transfection of eGFP dsRNA, we were able to show that the 22nt siRNA response is a property of the host cells, not the virus. Ours was one of the first studies to characterise the antiviral RNAi response in ticks.

In the final paper in this section, our attention turns to piRNAs [43]. As we have stated above, piRNAs are produced by a different pathway from siRNAs, independent of Dicer. They are longer, and have a broader size range: 25-29nt. Wu *et al* were the first to suggest that the piRNA pathway has an antiviral function [35], and subsequently piRNA signatures have been discovered in response to many other viruses, including BTV and SBV (above: [38]). Our study aimed to test that hypothesis that PIWI proteins are involved in the antiviral response in mosquito cells: if they are involved, then it would be expected that silencing the proteins involved would increase viral output. By infecting mosquito cells with Semliki forest virus (SFV) and selectively silencing various PIWI proteins, we were able to measure the effect of PIWI protein knock downs on SFV replication and indeed the piRNA response. Knockdown of certain PIWI proteins resulted in a lower piRNA response and higher SFV replication, suggesting that they are involved in the antiviral response.

3. Salmonella species as important pathogens of farm animals

The species *Salmonella enterica* includes over 2600 serovars representing hugely important pathogens of both animals and humans. As farm animal pathogens, *Salmonella* serovars enter the human food chain, increasing the chances of zoonosis. By definition, zoonotic pathogens maintain the ability to infect different host species (host generalists); however, within the *Salmonella* serovars, there are many which are only able to infect a single host (host specificists). Much of the genomic research in *Salmonella* has exploited these differences by looking for signatures of host adaption encoded within the genomes of *Salmonella* serovars.

3.1 Sequencing *Salmonella* genomes

In 2008, I was part of a group who sequenced, annotated and published the genomes of two *Salmonella enterica* serovars [44]: Enteritidis PT4, a host generalist; and Gallinarum 287/91, a serovar that only infects poultry. *S.* Enteritidis is now regarded as a pathogen of major public health significance, causing human food poisoning in many areas of the World [45]. Infection with *S.* Enteritidis results in fever, abdominal pain and diarrhoea. By contrast, *S.* Gallinarum predominately infects poultry, and is the causative agent of fowl typhoid [46]. *S.* Gallinarum also tends to cause systemic disease, and therefore host adaptation has also co-evolved with a change in habitat within the host. It has been proposed previously that *S.* Gallinarum and *S.* Enteritidis shared a common ancestor [47], with *S.* Gallinarum becoming non-motile due to a mutation in the *fliC* gene [48].

Our study pre-dated next-generation sequencing, therefore both genomes were sequenced using traditional Sanger shotgun sequencing and annotated manually and through comparison to *S.* Typhimurium LT2. Many similarities between Enteritidis PT4 and Typhimurium LT2 were found, including presence of pathogenicity islands and fimbrial gene clusters, a feature consistent in other promiscuous salmonellae [49]. However, the most striking output from the study was that, firstly, the genome sequences of the Gallinarum and Enteritidis strains showed a huge amount of sequence similarity, consistent with the hypothesis of a common ancestor; and secondly, the Gallinarum genome harbouring a large number of pseudogenes compared to both Enteritidis PT4 and Typhimurium LT2. Over 300 genes carried predicted frameshifts or stop codons, representing about 7% of the total coding capacity of the genome. Important pathways and functions that have been lost include metabolic pathways, restricting the available sources of carbon and energy,

including propanediol metabolism. Mutations also exist in the glycogen biosynthesis pathway and in amino acid catabolisis and biosynthesis. Finally, mutations exist in genes related to motility and chemotaxis, consistent with the non-motile nature of Gallinarum. It is impossible to assign an evolutionary role to all of these mutations, however, the gene loss is striking and we hypothesised that gene loss in metabolic pathways restricts the ability of Gallinarum to survive outside of the host and contributes significantly to host adaptation. Gene loss related to fimbrae and flagella may be methods by which Gallinarum evades the host immune system. We suggested that more research is needed to fully investigate the contribution of these gene losses to host adaptation.

Some of this work was followed up in 2010/11 by a PhD student who I supervised with Mark Stevens and David Gally [50, 51]. This time using Illumina next-generation sequencing, we sequenced, assembled and annotated 4 *Salmonella* genomes of well-defined virulence in farm animals. These were *S.* Typhimurium ST4/74, originally isolated from a calf; *S.* Choleraesuis SCA50, originally isolated from a pig; *S.* Dublin SD3246, again originally isolated from a calf; and *S.* Gallinarum SG9, first isolated from a case of fowl typhoid. One striking conclusion from this work was that the actual sequencing component was trivial and cheap, taking only a few weeks and costing a few thousand pounds; compared to the bioinformatics analysis and (assembly, annotation) and data release, which took in excess of 12 months. This contrast inspired us to write a review of bacterial genome annotation, which I will discuss in more detail later [52]. Genome annotation of the new Gallinarum strain, SG9, revealed a 277 predicted pseudogenes, and a large overlap in the affected pathways between SG9 and 287/91. This adds further evidence that gene loss is an important method of host adaptation in *Salmonella* species.

3.2 A closer look at fimbrial operons

A further follow up study was carried out by another PhD student whom I co-supervised with Mark Stevens [53]. In this study we examined the repertoire, organisation and sequence of fimbrial operons in a range of *S.* Enteritidis serovars. We attempted to correlate the presence/absence of fimbrial operons with host specificity, and selectively mutated each operon before examining the effect on colonisation. A total of 14 fimbrial operons were identified, 13 genomic and one plasmid-located. Pseudogene-causing mutations were found to be enriched relative to the genomic mean in fimbrial operons of host-restricted serovars, implying a selection pressure to “lose” these genes. In contrast,

broad host range serovars appeared to have mostly intact fimbrial operons. However, the presence/absence of any single fimbrial operon could not be correlated with host specificity. We attempted to mutate each of the 13 chromosomally located fimbrial operons in *S.* Enteritidis, and then screened each of the mutants in a chick colonisation model, and only two of the fimbrial operons were found to be statistically significant. The *stbA:cat* mutant was recovered at lower levels than wild-type, as was the *peg:cat* mutant. This is particularly significant, as the *peg* operon was identified as a novel fimbrial operon in [44], displaying only 70% identity to the operon in the orthologous position in *S.* Typhimurium. The fact that the majority of fimbrial operons don't seem to affect colonisation in isolation may be due to compensatory effects. Indeed, it is known that gene expression of fimbrial operons is highly variable, and it is possible that *Salmonella* species switch these operons on and off as a method of evading the host immune system.

3.3 *Salmonella* transcriptomics during colonisation

The major mode of entry into the human food chain is through colonisation of the chicken caeca. *Salmonella enterica* serovars Typhimurium and Enteritidis are major human pathogens, and healthy adult chickens often show little or no symptoms after infection with these serovars. Furthermore, infection of chicks only a few days old results in caecal colonisation and persistent shedding, which contributes to carcass contamination at slaughter. In 2011, we published a study which used microarrays to compare RNA isolated from *S.* Typhimurium infected chicken caeca with RNA from *S.* Typhimurium grown in LB broth, aiming to identify genes involved in chicken caecal colonisation[54]. The results demonstrated a large amount of transcriptional changes between the samples, and demonstrated decreased metabolic activity in *Salmonella* growing in the caecal lumen compared to *Salmonella* growing in LB broth. Genes involved with the cell cycle and DNA replication were down-regulated in the caeca, suggesting that the bacterial cells are growing more slowly than in broth. Genes involved in flagellum production were down-regulated in the caecal lumen, as were several chemotaxis genes, and these may indicate an effort to evade the host immune system. Some fimbrial genes were up-regulated, which may represent phased expression of some of the fimbrial operons, again potentially to evade the immune response. A large number of metabolic genes were differentially expressed, which is not surprising given the vastly different environments, energy and oxygen sources represented by the caecal lumen and LB broth.

3.4 Reviewing bacterial genome annotation

I first began analysing bacterial transcriptomic data in 2002, and began to include analysis of bacterial genomic data in 2004. It became very clear very quickly that all of these genomic and post-genomic approaches relied very heavily on two things: firstly, a high quality reference genome, and secondly, a high quality annotation of that reference. After our experiences assembling, annotating and releasing 4 *Salmonella* genomes in 2011, we were inspired to write a very well received review on the automatic annotation of bacterial genomes [52]. An important driver for this is the pace of change in bacterial genomics and DNA sequencing. In 2008, it cost several hundred thousand pounds to sequence and assemble two *Salmonella* genomes. Whilst automatic annotation occurred, each gene was checked by eye using the Artemis software, and comparisons between genomes were also carried out by eye using the Artemis Comparison Tool. I recall teaching our PhD students how to use ACT, and sitting with them as we visually searched for pathogenicity islands and regions of difference (RODs). Today, a bacterial genome can be sequenced for a few hundred pounds [55]. The impact of this is huge, as annotation efforts move away from manual to automatic. The purpose of our review was to highlight major areas of concern in the automatic annotation process, and highlight areas that could be improved. A major issue that we highlighted is the propagation of errors – as most automatic genome annotation relies on transference of annotation between homologous sequences, if the reference genome annotation is incorrect, this will be propagated. Using the example of the *eutN/eutM* locus in *Salmonella* species, we demonstrated that it is impossible to resolve this locus in certain genomes using sequence evidence. Alternately annotated as a single intact gene, a single pseudogene, two intact genes, two pseudogenes and one intact gene or one pseudogene, the *eutN/eutM* locus in *Salmonella* is an unsolvable problem for automatic annotation where only sequence evidence is used. Further annotation of this locus in novel *Salmonella* genomes will be defined not by biology, but by the choice of reference genome used. We also identified orthologous genes that should have the same gene name but do not – a subset of the nomenclature problem we highlighted further in the review, where we discovered 23,843 sets of genes that had differing product/protein names. The gene *tnp* was identified as the worst, having a total of 151 different product/protein names across all genomes. We also highlight the issue of spelling mistakes, using as an example “syntase”, a mis-spelling of “synthase” (as I wrote that, Microsoft Word auto-corrected “syntase” to “synthase”, proving that technology *can* solve

some of these issues). At the time of publication, 128 proteins in UniProt [56] contain the mis-spelled form “syntase”. This is not a protein, or indeed a UniProt problem, it is a genome annotation problem which has propagated through the public databases. In an attempt to address many of these problems, we proposed three measures. Firstly, a curated set of high quality, gold standard genomes which should be used as references. We observed that of 1851 published and completed bacterial genomes, only 102 had a version number of 0.2 or higher – so only 5% of submissions have been revised since publication. The genome of *Salmonella* Typhimurium LT2 was published in 2001, yet the sequence version number is 0.1. So the primary source of genome annotation for this strain has not been updated in over 13 years. Defining a small set of manually curated genomes with updated, high quality annotations may help us avoid some of these pitfalls. We also propose the implementation of auto-correcting software for improved genome annotation – we have already seen that MS Word can correct mis-spelled biological names, and Google Search can also do this. If at the point we transferred gene annotation, we also corrected it for common spelling mistakes, much progress could be made. Finally, we propose integration of new data types – RNA-Seq data, for example, could be used to resolve regions such as the *eutN/eutM* locus described previously.

4. Bioinformatics software development and application

Bioinformatics is a fascinating and innovative science which, in many ways, is no different to other branches of biology: we start with a problem, a question that needs to be answered; we develop a method that helps answer that question; if the method itself is innovative enough and useful to others, we publish that method; and we apply the method to a set of biological data, interpret the results and publish. However, it is not as young a science as many think, many recognise Margaret Dayhoff's "Comprotein" as the first bioinformatics publication [57], which appeared in 1962. I myself have been involved in bioinformatics since 1997 (however only entering academia in 2002). I joined the discipline in the pre-genome era, dealing with gene expression array data – arrays which were built from normalised cDNA libraries, as we didn't have a genome to work from. Since then, I have enjoyed three waves of "big data", each larger than the last, and each causing scientists to claim that data analysis will be the next big bottle neck (for example, see [58]). At first the task was to deal with the huge amounts of data from just a few genomes (human; mouse; rat); post-genome, the task was to deal with the huge amounts of microarray and proteomic data being generated; and in the next-generation sequencing era, the task is to deal with the large amount of sequencing data generated for each experiment. Of course, these are vastly different data types, and bioinformaticians need to have the ability to adapt to these data types.

Over the years, I have written, released, published and supported 7 software packages, all of which could be considered a novel "method" that enables biological research. Each was born from a problem I and my collaborators had encountered, for which a suitable solution was not available. Six of these software tools are packages for the statistical package R [59], which has had a remarkable impact on the biological sciences. The Bioconductor project [5], which currently contains over 800 packages designed for the analysis of biological data, has over 6000 citations according to Google Scholar.

4.1 Visualisation of quantitative data with bacterial genomes

My first software publication came in 2005 [60], and attempted to deal with the issue of how to visualise large amounts of quantitative data alongside genome structure and annotation information in bacteria. A particular driver was to enable this within R – we were analysing microarray data from bacteria using the limma [61] package, therefore having an integrated visualisation tool would be hugely beneficial. The visualisation of

quantitative data alongside traditional genome annotation is now common, and is often carried out by attaching tracks to genome browsers, or using tools such as IGV, IGB or Tablet. However, in 2005 this idea was in its infancy. Some software packages did exist to overlay numerical data on bacterial genomes, but many of these used a colour scale to represent the values, which didn't provide the granularity we required. Others produced genome-scale, circular visualisations which, whilst very pretty and attractive, were hard to interpret. Many of the tools were also only able to show a single value per gene, which is not useful for time-course experiments. ProGenExpress therefore found its niche in being able to produce high-quality, quantitative plots of numerical data integrated with genome annotation information and represented as a horizontal or vertical track. ProGenExpress could view the whole genome, or zoom into a particular region.

Bacterial genomes present particular issues which make visualisation interesting – genes are often co-expressed through operons, and there are important mobile elements such as prophage and pathogenicity islands that represent groups of genes that may be considered together, as a functional unit, rather than as separate genes. Presenting scientists with a list of differentially expressed genes, as is the result of limma, ignores this relatedness. An example is given in the paper – using a public time-course data set, visually one can see that all 14 genes of the *fli* operon in *Salmonella* are down-regulated in murine-macrophages relative to control (we have seen above that *Salmonella* species often reduce motility during infection!). However, 3 out of the 14 genes are not significantly down-regulated and would not appear in a simple gene list. Only by visualising those three genes in the context of the operon can we conclude that they are down-regulated too.

4.2 Software for pathogen detection microarrays

Microarrays weren't only being used for gene expression studies, they were also (and continue to be) used for pathogen detection. The simple idea is that if one prints probes representing a majority of known pathogens (viral, bacterial, funghal, parasitic etc) then one can take a clinical/infected sample, extract RNA and wash it over the microarray, and spots representing any pathogens present should light up. In reality, it is never as simple as this, however there were some notable successes. Many credit Joe DeRisi's lab with pioneering this work [62], though other groups had published previously [63], and the major advance Wang *et al* provided was the ability to use microarrays to discover *novel* viruses (rather than just those printed on the array). I became involved in a project funded

by Defra in 2005 to create a "Biochip" that would represent all of the major human, animal and plant viruses of interest to Defra, with the vision that we would be able to create a single diagnostic array that would be used across all Defra's disease surveillance laboratories. Being familiar with microarray data, my role was to ensure that we had robust software with which to analyse the data. The DeRisis group had recently published some software for pathogen-detection array data called E-Predict [64], however this software presented a problem for us. E-Predict worked by creating a model of theoretical hybridisation profiles based on BLAST similarity profiles of the probes on the array to public viral genomes. P-values were assigned to real samples by comparing the real data with the theoretical data using a similarity measure. What this meant was that E-Predict was intimately tied to the DeRisi array, and we would have to completely rebuild the model for our array in order to use the software. Also, E-Predict would need to be retrained every time a new set of probes was added to the array. For these reasons, I decided to develop DetectiV [65], a package for R that would be capable of analysing any set of pathogen detection microarray data.

DetectiV takes two sets of information – a set of numerical values extracted from an array and linked to probe IDs, and a set of phylogenetic information for the pathogen species represented by each probe. DetectiV recognised that such data would be noisy – extract RNA from an infected tongue lesion, and there would be RNA from the host, pathogenic and non-pathogenic organisms. Add to this the fact that RNA would cross hybridise both specifically and non-specifically, and one can anticipate a noisy profile. However, DetectiV relied on the simple assumption that the "real" signal would be stronger than the noise (this is an assumption of all bioinformatics software!). To aid in pathogen detection, we implemented three normalisation strategies. The first, borrowed from gene expression microarrays, calculated log2 of each measurement divided by the global median for all probes – here the global median represented the "noise". The second defined a set of control probes on the array, and calculated log2 of each measurement divided by the mean value of the control probes – here the control probes represented the noise. The final method defined an entire array as a control array, and calculated the log2 value of each measurement divided by its equivalent measurement on the control array. In all instances, "real" signal should be greater than zero. DetectiV then applied a t-test for each phylogenetic group, be that at the species, genus, family etc level testing the null hypothesis that the set of measurements for each group was no different from zero.

In publishing DetectiV, it was important to compare to the "state-of-the-art", E-Predict. We therefore used the data released by Urisman *et al* which was used to train and run E-Predict. Urisman *et al* had used E-Predict, and their pathogen detection array, to detect SARS, even though SARS was not represented by any probes on the array. We also decided to include a SARS analysis in the DeteciV results.

DetectiV proved more successful than E-Predict on their own dataset, successfully predicting the correct virus in 55 out of 56 arrays, whereas E-Predict was only successful in 53 out of 56. In addition, DetectiV was able to detect SARS by defining a set of probes that showed > 80% similarity to the SARS genome. We then validated DetectiV on a completely new set of data – 8 arrays from different foot-and-mouth-disease (FMDV) types, and a further 4 arrays from Avian infectious bronchitis virus (IBV). In all 12 cases, DetectiV predicted the correct result. We were unable to test E-predict on the new data, as this would have entailed re-training of the entire model, and in many ways this shows the power of DetectiV – its simplicity enabled users of vastly different arrays to use the software without any need for complex re-training. Despite being published in 2007, I know that DetectiV is still in use in diagnostic labs today.

4.3 Metagenomics: dealing with data from multiple genomes

The term metagenomics refers to the study of all genomes within an ecosystem, and has only truly been enabled by the ultra-high-throughput nature of second and third generation sequencing. Metagenomics is one technique by which researchers can study the microbiome, which itself refers to the entire complement of microbes (bacteria, archaea, protists, funghi, viruses) that live in a particular ecosystem. The diversity and novelty of organisms in every ecosystem is huge, with organisms adapted to all types of environment, including extreme conditions. Metagenomic approaches increase the "sequencing space" from which we can discover novel biocatalysts [66, 67]. In 2010, we were awarded funding from the Technology Strategy Board to do just that – sequence a new environment, the rumen gut microbiome, in an attempt to discover new enzymes involved in the breakdown of cellulose.

Second generation sequencing experiments routinely produce many hundreds of gigabases of sequence data, and therefore offer a unique insight into the genomes of microorganisms living in an environment. In order to discover novel genes and enzymes, researchers must reconstruct genomes into sufficiently large fragments to allow full genes to be predicted.

Once gene predictions are complete, predicted protein sequences can be created by translating the genomic sequence, and protein domains assigned. Protein domain assignments using, for example, profile hidden markov models (HMMs) are more sensitive to distant matches than homology searches such as BLAST, and are therefore more likely to be useful in a metagenomic context. Our studies in ruminants suggest that over 90% of such metagenomic gene predictions do not have a match in NR at over 90% identity. These results are currently unpublished, but they are similar to other published findings - Venter *et al.* [68] reported over 1.2 million novel genes, and Hess *et al.* [69] reported over 2.5 million putative genes, 27755 containing a domain relevant to biomass degradation.

The term "assembly" refers to the process by which researchers reconstruct the genome of an organism from sequenced fragments, and there have been several paradigms suggested for doing this, including overlap-layout-consensus, De Bruijn graphs and string graphs (reviewed in [70]). Given that assembling a single genome completely is still an active area of research, attempts to simultaneously assembly the multiple genomes within a metaganomics experiment are incredibly difficult. One issue is that metagenomic assembly graphs are larger and require more memory and compute resources. Ray Meta [71] makes use of distributed computing and message passing; Pell *et al* [72] use a bloom filter and kmer connectivity to partition the graph. Meta IDBA [73] also uses connectivity to reduce the complexity of graph, whereas MetaVelvet [74] uses both coverage binning and connectivity.

Once an assembly is complete, annotation of the contigs and scaffolds must take place. There are additional problems that can be encountered here over and above the problems encountered during traditional bacterial genome annotation (which we reviewed in [52]). A number of tools have been published which focus specifically on metagenomic gene prediction, including MetaGeneAnnotator [75], Orphelia [76], FragGeneScan [77], and Glimmer-MG [78]. Once we know the putative location of genes, protein-coding genes can be further annotated with known domains, using tools such as HMMER [79] and InterProScan [80] and databases such as Pfam [81] and InterPro [82].

In 2012, we were in exactly this position. Having deep sequenced DNA extracted from the rumen of 12 different animals, we had several hundred thousand assembled scaffolds, over 3.7 million predicted genes and proteins with over 1.4 million domains annotated upon them. Our task was to find novel enzymes involved in the breakdown of cellulose. In

biological research, it is hugely important to form collaborative teams between bioinformaticians and the biologists with the expert knowledge to interpret the findings of *in silico* analyses. In this instance, the experts were biochemists and experts in ruminal nutrition. The problem very rapidly became: how do we share the results of our metagenomic assembly and annotation with our collaborators?

In 2012/13, we designed, built, released and published Meta4, a simple yet powerful web application that allows the easy sharing of metagenomic assembly and annotation results with collaborators who do not have expertise in bioinformatics[83]. Meta4 fulfilled several unmet requirements in the field of metagenomics. Unlike tools such as IGM/M [84], CAMERA [85] and MG-RAST [86], which focus on complete annotation and comparison problems, Meta4 is specifically focused on allowing researchers to search large and complex datasets for novel versions of existing/known enzymes. At the heart of this search is the annotation of known domains – therefore users of Meta4 can search for protein/gene sequences that have a particular domain and length. Picking a domain such as "PF00150 Cellulase", for example, will bring up a list of all proteins in the database that contain a cellulase domain.

Meta4 is organised as a simple relational database, in MySQL, with scripts written in Perl that allow users to upload data to the database from common file formats such as FASTA and GFF. Further Perl scripts, using the common gateway interface (CGI) of an Apache web server, provide a user-friendly front-end that enables researchers to query the database. Meta4 is not intended to be a large global database of all known metagenomic datasets; rather, we envisaged that groups would set up a new instance of Meta4 for each new dataset, with the specific aim of sharing data with collaborators. It is possible to set up a new instance of Meta4 in less than an hour, and an Amazon Machine Image (AMI) is available on Amazon EC2 that will allow users to create a cloud based instance.

A feature of Meta4 is its use of web services. By necessity, the database stores protein domain annotations so that users are able to query by domain; however, when looking at the gene prediction page, Meta4 uses the InterProScan and EBI BLAST web services to serve up the latest information for the protein prediction in question, ensuring users gain access to the most recent searches for that protein.

Meta4 has now been reliably working in my group for over 18 months, sharing data from our metagenomic sequencing and bioinformatics effort with Ingenza ltd, a small biotech company focused on the development of technologies for biofuels. Using Meta4, they were able to take the 3.8 million proteins and reduce that dataset to around 100 candidate enzymes. We are also using Meta4 to share data from a separate experiment involving methane emissions.

4.4 A change in paradigm: nanopore sequencing data

Nanopore sequencing represents a paradigm shift in DNA sequencing, and today it is the only sequencing technology that measures an actual single molecule of DNA, rather than incorporation events into a template strand. In nanopore sequencing, a protein nanopore is attached to a membrane and changes in electronic signal are measured as single molecules of DNA pass through the pore. Although research in solid-state nanopores is ongoing, most success has been had so far using biological nanopores such as α-hemolysin [87-89] or MspA [90]. Much of the innovative research in this area has been carried out at the University of Oxford by Prof Hagan Bayley's group, and it was Prof Bayley who formed Oxford NanoLabs, now Oxford Nanopore Technologies (ONT). ONT are the first company to bring to market a commercial nanopore sequencing device. In the field of what ONT term "strand sequencing", they have various products: MinION, GridION and PromethION. However, only one is now available to researchers, the MinION.

The MinION is the world's first mobile DNA sequencing device, measuring 10cm in length and powered by the USB port of a laptop. The MinION device is compatible with consumable flowcells, each containing the sensor chip, application specific integrated circuit (ASIC) and nanopores needed to perform single-molecule sequencing. MinION flowcells have 512 channels, each designed to hold a single engineered protein nanopore. Nanopores are embedded within a membrane and an ionic current is passed through the nanopore by setting a voltage across this membrane. When a DNA molecule passes through the nanopore, this disrupts the current in a reproducible fashion, allowing the measurement of the set of bases that occupy the aperture at that exact moment. As many bases occupy the aperture at any one time, ONT have developed a proprietary base-calling algorithm based on hidden-markov models (HMMs).

4.4.1 MinION workflow

At present, the MinION library preparation workflow takes approximately 4 hours, and involves shearing of the DNA, end-repair, clean up and ligation before a sample can be pipetted onto the MinION. These protocols are under active development, and several updates have been released since the beginning of the MinION access programme (MAP).

Each MinION has a dedicated, high-specification laptop running Windows 7, with 8Gb RAM and a 128Gb solid state drive. The MinION device is managed by the MinKNOW software which gives details on pore occupancy, read length and throughput during the run. As soon as the library is placed onto the device, the MinION begins sequencing. Each channel/nanopore reports asynchronously, creating a single file per channel per read. These are created in HDF5, a compressed binary hierarchical data format. Surprisingly, there are no run folders, and all data files are written to a single directory (e.g. C:\MinION). This rapidly becomes confusing as files from multiple runs exist in the same directory. At that time, the HDF files contain metadata about the run as well as information on the signal and events as a DNA molecule passed through the nanopore. This single folder is monitored by an agent called "metrichor", which uploads new files to a cloud-based base-caller. Once base-called, the files are downloaded from the cloud, again into a single directory (e.g. C:\MinION\downloads). At this stage the HDF5 files contain additional information, the addition of sequence data in fastq format. A key feature of ONT's technology is "2D" reads. During library prep, DNA molecules have a hairpin adapter ligated to them, meaning the molecule can pass through the pore once, round the hairpin, and pass through a second time. Therefore, each DNA molecule may be read twice. These reads are called 2D reads and are of a higher quality than 1D reads.

At this early stage, the requirements for library preparation, and reliance on a cloud-based base-caller, limit the mobility of the MinION device; however, it is clear that the future of the MinION is mobile DNA sequencing, with expected advances in sample preparation, and a stand-alone base-caller in the pipeline. ONT have promised that a standalone, non cloud-based base-caller will be released in the very near future. Advances in library preparation are also expected that will allow field-based sequencing of samples. It is essential that bioinformatics tools are developed that allow users of the MinION, both in the lab and in the field, to query and analyse nanopore sequencing data.

4.4.2 poRe

We have developed, released and published the first version of poRe, an R package that enables researchers to work with ONT MinION data [91]. At time of publication, only one other software tool was available that helped researchers work with MinION data, poretools [92]. Poretools is a python library, and was published approx. one week before poRe. Crucially, poRe has several advantages over poretools, including ease of installation and availability within R [59], a sophisticated mathematical and statistical environment which freely available on Windows, Linux and Mac.

poRe depends upon the RDHF5 package, an R interface to the HDF5 library. HDF5 is a hierarchical, binary data format that allows the construction of arbitrary hierarchical data and paths. The hierarchical format allows random access of particular data sets by providing the hierarchical path to that data within the file (rather than having to read the whole file). Therefore very large files can be queried quickly. However, the HDF5 offers only low-level C functions to access the data, and RHDF5 simply wraps those low-level functions in R. Therefore, within poRe, we have written some higher level functions that allow users to query and extract data from the MinION HDF5 files, and we can do this because we know the specific path to particular datasets within those files. An example of the FAST5 file structure can be seen in the supplementary data to our publication [91].

poRe allows users to perform several essential functions. The first is simply organisation. The MinION writes data from multiple runs into the same folder, and users are often faced with 10,000s of FAST5 files and no way to easily interrogate or analyse them. However, embedded within each FAST5 file are two key pieces of information: firstly, a random text string which uniquely identifies a run; and secondly, the name and version of the cloud base-caller used to extract FASTQ data. poRe will read every single FAST5 file within a specified folder, and copy them to new folders based on the run name and the name and software version of the base-caller.

Once files and data are organised, poRe also allows extraction of FASTQ and FASTA data, collection of key metrics about the run, plots of yield and read length, and extraction of the raw events data. The raw events data can also be plotted, as a “squiggle” plot. poRe received quite a lot of attention whilst the paper was a pre-print on bioRxiv, and has several users across Europe and in the USA.

4.5 MicroRNAs and gene-set enrichment

During our research into Marek's disease virus, questions about the function and targets of both host and pathogen microRNAs naturally arose. MicroRNAs are small 21-23 nucleotide RNA molecules that act as guides for the RNA-induced silencing complex (RISC), and which mediate the decay of target mRNA molecules. MicroRNAs therefore act as negative regulatory mechanism by preventing the translation of mRNA molecules into proteins. Despite extensive research, the exact mechanism by which microRNAs target mRNAs is unknown. MicroRNAs bind, via sequence complementarity, to target mRNAs, with binding sites being found most often in the 3' UTR. However, sequence complementarity very rarely exists along the entire microRNA, with the "seed" region (positions 2-7) being particularly important. This is an ongoing area of research, and published algorithms include TargetScan [93], miRanda [94], RNAHybrid [95] and PicTar [96].

Utilising the results of these target prediction algorithms in larger experiments became important. Specifically, there was a requirement to link the outputs of mRNA and miRNA expression studies. For example, one might somehow extract a list of down-regulated genes between a pair of conditions, and ask the question "which microRNAs are implicated in the regulation of these genes?". This is exactly the question that CORNA was designed to answer, a piece of software we wrote and published in 2009 [97]. CORNA was one of the first software tools, along with SIGTERMS [98], that allowed users to input a list of mRNA genes and a list of predicted miRNA-mRNA targets, and which would then predict which microRNAs might be regulating the input gene list. CORNA had several advantages over SIGTERMS, the major one being that CORNA is implemented in R and can therefore integrate with existing bioinformatics pipelines; whereas SIGTERMS is implemented in Excel, a Windows only platform that is rarely used in bioinformatics.

CORNA uses the hypergeometric distribution to look for enriched microRNA target predictions in a user supplied list of mRNA genes. For each microRNA, an observed number of miRNA targets in the sample provided is compared to an expected number of microRNA targets calculated from a population, in this case the complete set of predicted miRNA-mRNA relationships for the species in question. P-values can then be calculated using the hypergeometric distribution, and corrected (if necessary) using a variety of adjustment methods. The users is therefore supplied with a list of microRNAs, the observed and expected number of targets in their gene list, and a p-value for the enrichment of that

microRNA in their dataset. Whilst we demonstrated and published CORNA on a simulated dataset, we had already used the software for our own data, and achieved impressive results. These are as yet unpublished, though we have a draft paper and hope to submit soon. To summarise, using microarrays we established a set of genes in chicken embryo fibroblasts whose expression was down-regulated over a time-course of infection with REV-T (Reticuloendotheliosis virus strain T), a virus that causes B- and T- cell lymphomas in chickens. When that gene list is submitted to CORNA, along with the predicted miRNA-mRNA targets predicted by miRanda, the top hit is gga-miR-155, a microRNA known to be involved in oncogenic pathways and regulation of the immune response. Lab work confirmed that over the same time-course, gga-miR-155 increases in expression massively. CORNA also predicted several of the miR-17-92 cluster, a known cluster of oncomirs (reviewed in [99]). CORNA is therefore capable of predicting the likely involvement of microRNAs in the regulation of gene lists supplied to it. I was part of a group that went on to review the field of miRNA-mRNA bioinformatics published in 2010 [100].

Soon after the publication of the CORNA paper, I was involved in two additional publications relating to the zebra finch, a model organism often used to research the evolution of learned speech. The genome of *Taeniopygia guttata* was sequenced, assembled, annotated and published in 2010 [101]. During this project, the expression of genes up- and down- regulated after exposure to song was measured, as were lists of genes under positive selection. These gene lists were submitted to CORNA to measure the enrichment of GO terms, and it was CORNA that identified the enrichment of ion channel genes, which are known to have functions in neurological function. CORNA was therefore central to the paper linking ion channel genes to song behaviour.

In addition to the genome paper, I was also part of a group who investigated the expression of microRNAs before and after song exposure in the forebrain of zebra finches [102]. This work involved the analysis of Illumina next-generation sequencing data, and allowed us to identify both known and novel microRNAs whose expression changes in response to song.

4.6 Analysis of small RNAs in insect vectors of viral disease

I began collaborating with Alain Kohl's group in 2011, and we have published several papers since then, some of them detailed in section 2. The goal in each case was to analyse large sequencing datasets to summarise, analyse and visualise the patterns of small RNA expression, investigating different viruses and how their hosts (insects and ticks) reacted to

them. Illumina sequencing datasets produce millions of short reads, and it became very quickly apparent that the software tools available were inadequate, inflexible and difficult to use. Both Paparazzi [103] and Visitor [104] are all-in-one "pipelines" that take sequencing data, run alignments (using a single aligner) and produce a range of plots as output. Both require use of the Linux operating system, something which many biologists are not familiar with. We wanted to have more control over the alignment stage, and also to be able to control the appearance of any plots.

We published viRome [105] in 2013. Crucially, viRome does not include the alignment stage, and allows researchers to choose and run their own alignment software; there are almost 100 to choose from [106], each of which has parameters that can be optimised, and separating data analysis from sequence alignment allows far more flexibility. ViRome reads in data in the BAM format, a standard format for sequence alignment. Using R as the platform allows researchers to access the huge and powerful range of statistical and mathematical functions available as part of the core platform and add-on packages. R is also multi-platform, running on Windows, Linux and Mac; therefore by using viRome, researchers instantly have more power and flexibility than if they were to use Paparazzi or Visitor.

ViRome allows users to plot read length distributions on different reference genomes, which can be used as evidence for a siRNA (21-22nt) or piRNA (25-29nt) response. The position and strand of these alignments can also be visualised along the length of the genome sequence, to look for bias and hot-spots. Other patterns of piRNA expression are a U_1 and A_{10} bias, and a peak of 10 nucleotides when looking at the frequency of gaps between 5' ends of reads mapped to opposite strands, due to "ping-pong" amplification. These can also be analysed and visualised using viRome. The software can also summarise all alignments and output these to a CSV file that can be opened in Excel.

ViRome is a very flexible package and enables users a far greater flexibility in their data analysis. The software was actually used in several publications prior to release and publication, therefore should have garnered more citations. We continue to use viRome in our research today and I expect several more publications in 2015

5. References

1. International Chicken Genome Sequencing C: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004, 432(7018):695-716.
2. Bovine Genome S, Analysis C, Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE *et al*: The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 2009, 324(5926):522-528.
3. Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ *et al*: Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 2012, 491(7424):393-398.
4. Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W *et al*: The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* 2014, 344(6188):1168-1173.
5. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* 2004, 5(10):R80.
6. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 1998, 391(6669):806-811.
7. Rocheleau CE, Downs WD, Lin R, Wittmann C, Bei Y, Cha YH, Ali M, Priess JR, Mello CC: Wnt signaling and an APC-related gene specify endoderm in early C. elegans embryos. *Cell* 1997, 90(4):707-716.
8. Fagard M, Boutet S, Morel JB, Bellini C, Vaucheret H: AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proceedings of the National Academy of Sciences of the United States of America* 2000, 97(21):11650-11654.
9. Catalanotto C, Azzalin G, Macino G, Cogoni C: Gene silencing in worms and fungi. *Nature* 2000, 404(6775):245.
10. Hamilton AJ, Baulcombe DC: A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* 1999, 286(5441):950-952.
11. Zamore PD, Tuschl T, Sharp PA, Bartel DP: RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* 2000, 101(1):25-33.
12. Lee RC, Feinbaum RL, Ambros V: The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993, 75(5):843-854.
13. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 2000, 403(6772):901-906.
14. Hammond SM, Bernstein E, Beach D, Hannon GJ: An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature* 2000, 404(6775):293-296.
15. Lupiani B, Lee LF, Cui X, Gimeno I, Anderson A, Morgan RW, Silva RF, Witter RL, Kung HJ, Reddy SM: Marek's disease virus-encoded Meq gene is involved in transformation of lymphocytes but is dispensable for replication. *Proceedings of the National Academy of Sciences of the United States of America* 2004, 101(32):11815-11820.

16. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* 2012, 7(3):562-578.
17. Jansson MD, Lund AH: MicroRNA and cancer. *Molecular oncology* 2012, 6(6):590-610.
18. Yao Y, Zhao Y, Xu H, Smith LP, Lawrie CH, Watson M, Nair V: MicroRNA profile of Marek's disease virus-transformed T-cell line MSB-1: predominance of virus-encoded microRNAs. *Journal of virology* 2008, 82(8):4007-4015.
19. Akiyama Y, Kato S: Two cell lines from lymphomas of Marek's disease. *Biken journal* 1974, 17(3):105-116.
20. Burnside J, Morgan RW: Genomics and Marek's disease virus. *Cytogenetic and genome research* 2007, 117(1-4):376-387.
21. Yao Y, Zhao Y, Smith LP, Lawrie CH, Saunders NJ, Watson M, Nair V: Differential expression of microRNAs in Marek's disease virus-transformed T-lymphoma cell lines. *The Journal of general virology* 2009, 90(Pt 7):1551-1559.
22. Lind EF, Ohashi PS: Mir-155, a central modulator of T-cell responses. *European journal of immunology* 2014, 44(1):11-15.
23. Zhao Y, Yao Y, Xu H, Lambeth L, Smith LP, Kgosana L, Wang X, Nair V: A functional MicroRNA-155 ortholog encoded by the oncogenic Marek's disease virus. *Journal of virology* 2009, 83(1):489-492.
24. Yao Y, Zhao Y, Smith LP, Watson M, Nair V: Novel microRNAs (miRNAs) encoded by herpesvirus of Turkeys: evidence of miRNA evolution by duplication. *Journal of virology* 2009, 83(13):6969-6973.
25. Yao Y, Smith LP, Petherbridge L, Watson M, Nair V: Novel microRNAs encoded by duck enteritis virus. *The Journal of general virology* 2012, 93(Pt 7):1530-1536.
26. Yao Y, Charlesworth J, Nair V, Watson M: MicroRNA expression profiles in avian haemopoietic cells. *Frontiers in genetics* 2013, 4:153.
27. Yao Y, Smith LP, Nair V, Watson M: An avian retrovirus uses canonical expression and processing mechanisms to generate viral microRNA. *Journal of virology* 2014, 88(1):2-9.
28. Payne LN: Retrovirus-induced disease in poultry. *Poultry science* 1998, 77(8):1204-1212.
29. Bizub D, Katz RA, Skalka AM: Nucleotide sequence of noncoding regions in Rous-associated virus-2: comparisons delineate conserved regions important in replication and oncogenesis. *Journal of virology* 1984, 49(2):557-565.
30. Bai J, Payne LN, Skinner MA: HPRS-103 (exogenous avian leukosis virus, subgroup J) has an env gene related to those of endogenous elements EAV-0 and E51 and an E element found previously only in sarcoma viruses. *Journal of virology* 1995, 69(2):779-784.
31. Schwartz DE, Tizard R, Gilbert W: Nucleotide sequence of Rous sarcoma virus. *Cell* 1983, 32(3):853-869.
32. Ding SW, Lu R: Virus-derived siRNAs and piRNAs in immunity and pathogenesis. *Current opinion in virology* 2011, 1(6):533-544.
33. Siomi MC, Sato K, Pezic D, Aravin AA: PIWI-interacting small RNAs: the vanguard of genome defence. *Nature reviews Molecular cell biology* 2011, 12(4):246-258.
34. Kim VN, Han J, Siomi MC: Biogenesis of small RNAs in animals. *Nature reviews Molecular cell biology* 2009, 10(2):126-139.
35. Wu Q, Luo Y, Lu R, Lau N, Lai EC, Li WX, Ding SW: Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proceedings of*

the National Academy of Sciences of the United States of America **2010, 107(4):1606-1611.**

36. **Blair CD: Mosquito RNAi is the major innate immune pathway controlling arbovirus infection and transmission. *Future microbiology* 2011, 6(3):265-277.**
37. **Donald CL, Kohl A, Schnettler E: New Insights into Control of Arbovirus Replication and Spread by Insect RNA Interference Pathways. *Insects* 2012, 3(4):511-531.**
38. **Schnettler E, Ratinier M, Watson M, Shaw AE, McFarlane M, Varela M, Elliott RM, Palmarini M, Kohl A: RNA interference targets arbovirus replication in Culicoides cells. *Journal of virology* 2013, 87(5):2441-2454.**
39. **Beer M, Conraths FJ, van der Poel WH: 'Schmallenberg virus'--a novel orthobunyavirus emerging in Europe. *Epidemiology and infection* 2013, 141(1):1-8.**
40. **Schnettler E, Tykalova H, Watson M, Sharma M, Sterken MG, Obbard DJ, Lewis SH, McFarlane M, Bell-Sakyi L, Barry G *et al*: Induction and suppression of tick cell antiviral RNAi responses by tick-borne flaviviruses. *Nucleic acids research* 2014, 42(14):9436-9446.**
41. **Kemp C, Imler JL: Antiviral immunity in drosophila. *Current opinion in immunology* 2009, 21(1):3-9.**
42. **Kurscheid S, Lew-Tabor AE, Rodriguez Valle M, Bruyeres AG, Doogan VJ, Munderloh UG, Guerrero FD, Barrero RA, Bellgard MI: Evidence of a tick RNAi pathway by comparative genomics and reverse genetics screen of targets with known loss-of-function phenotypes in Drosophila. *BMC molecular biology* 2009, 10:26.**
43. **Schnettler E, Donald CL, Human S, Watson M, Siu RW, McFarlane M, Fazakerley JK, Kohl A, Fragkoudis R: Knockdown of piRNA pathway proteins results in enhanced Semliki Forest virus production in mosquito cells. *The Journal of general virology* 2013, 94(Pt 7):1680-1689.**
44. **Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, Quail MA, Stevens M, Jones MA, Watson M *et al*: Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome research* 2008, 18(10):1624-1637.**
45. **Rodrigue DC, Tauxe RV, Rowe B: International increase in Salmonella enteritidis: a new pandemic? *Epidemiology and infection* 1990, 105(1):21-27.**
46. **Shivaprasad HL: Fowl typhoid and pullorum disease. *Revue scientifique et technique* 2000, 19(2):405-424.**
47. **Li J, Smith NH, Nelson K, Crichton PB, Old DC, Whittam TS, Selander RK: Evolutionary origin and radiation of the avian-adapted non-motile salmonellae. *Journal of medical microbiology* 1993, 38(2):129-139.**
48. **Kilger G, Grimont PA: Differentiation of Salmonella phase 1 flagellar antigen types by restriction of the amplified fliC gene. *Journal of clinical microbiology* 1993, 31(5):1108-1110.**
49. **Townsend SM, Kramer NE, Edwards R, Baker S, Hamlin N, Simmonds M, Stevens K, Maloy S, Parkhill J, Dougan G *et al*: Salmonella enterica serovar Typhi possesses a unique repertoire of fimbrial gene sequences. *Infection and immunity* 2001, 69(5):2894-2901.**
50. **Richardson EJ: Next-generation bioinformatics analysis of bacterial genomes, with a focus on serovar host specificity and pathogenicity in Salmonella. University of Edinburgh; 2013.**

51. Richardson EJ, Limaye B, Inamdar H, Datta A, Manjari KS, Pullinger GD, Thomson NR, Joshi RR, Watson M, Stevens MP: Genome sequences of Salmonella enterica serovar typhimurium, Choleraesuis, Dublin, and Gallinarum strains of well-defined virulence in food-producing animals. *Journal of bacteriology* 2011, 193(12):3162-3163.
52. Richardson EJ, Watson M: The automatic annotation of bacterial genomes. *Briefings in bioinformatics* 2013, 14(1):1-12.
53. Clayton DJ, Bowen AJ, Hulme SD, Buckley AM, Deacon VL, Thomson NR, Barrow PA, Morgan E, Jones MA, Watson M *et al*: Analysis of the role of 13 major fimbrial subunits in colonisation of the chicken intestines by Salmonella enterica serovar Enteritidis reveals a role for a novel locus. *BMC microbiology* 2008, 8:228.
54. Harvey PC, Watson M, Hulme S, Jones MA, Lovell M, Berchieri A, Jr., Young J, Bumstead N, Barrow P: Salmonella enterica serovar typhimurium colonizing the lumen of the chicken intestine grows slowly and upregulates a unique set of virulence and metabolism genes. *Infection and immunity* 2011, 79(10):4105-4121.
55. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH, Phillippy AM: Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome biology* 2013, 14(9):R101.
56. UniProt C: Activities at the Universal Protein Resource (UniProt). *Nucleic acids research* 2014, 42(Database issue):D191-198.
57. Dayhoff MO, Ledley RS: Comprotein: a computer program to aid primary protein structure determination. In: *Proceedings of the December 4-6, 1962, fall joint computer conference.* Philadelphia, Pennsylvania: ACM; 1962: 262-274.
58. Experts Say Analysis Is The Next Big Crunch The Genomics Community Faces [http://www.genomeweb.com/informatics/experts-say-analysis-next-big-crunch-genomics-community-faces]
59. R Core Team: R: A Language and Environment for Statistical Computing. In. Vienna, Austria: R Foundation for Statistical Computing; 2014.
60. Watson M: ProGenExpress: visualization of quantitative data on prokaryotic genomes. *BMC bioinformatics* 2005, 6:98.
61. Smyth GK: limma: Linear Models for Microarray Data. 2005:397-420.
62. Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J *et al*: Viral discovery and sequence recovery using DNA microarrays. *PLoS biology* 2003, 1(2):E2.
63. Lapa S, Mikheev M, Shchelkunov S, Mikhailovich V, Sobolev A, Blinov V, Babkin I, Guskov A, Sokunova E, Zasedatelev A *et al*: Species-level identification of orthopoxviruses with an oligonucleotide microchip. *Journal of clinical microbiology* 2002, 40(3):753-757.
64. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, Wang D, DeRisi JL: E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome biology* 2005, 6(9):R78.
65. Watson M, Dukes J, Abu-Median AB, King DP, Britton P: DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. *Genome biology* 2007, 8(9):R190.
66. Cowan DA, Arslanoglu A, Burton SG, Baker GC, Cameron RA, Smith JJ, Meyer Q: Metagenomics, gene discovery and the ideal biocatalyst. *Biochemical Society transactions* 2004, 32(Pt 2):298-302.
67. Cowan D, Meyer Q, Stafford W, Muyanga S, Cameron R, Wittwer P: Metagenomic gene discovery: past, present and future. *Trends in biotechnology* 2005, 23(6):321-329.

68. **Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al*: Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004, 304(5667):66-74.**
69. **Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T *et al*: Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 2011, 331(6016):463-467.**
70. **Nagarajan N, Pop M: Sequence assembly demystified. *Nature reviews Genetics* 2013, 14(3):157-167.**
71. **Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J: Ray Meta: scalable de novo metagenome assembly and profiling. *Genome biology* 2012, 13(12):R122.**
72. **Pell J, Hintze A, Canino-Koning R, Howe A, Tiedje JM, Brown CT: Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences of the United States of America* 2012, 109(33):13272-13277.**
73. **Peng Y, Leung HC, Yiu SM, Chin FY: Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 2011, 27(13):i94-101.**
74. **Namiki T, Hachiya T, Tanaka H, Sakakibara Y: MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research* 2012, 40(20):e155.**
75. **Noguchi H, Taniguchi T, Itoh T: MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA research : an international journal for rapid publication of reports on genes and genomes* 2008, 15(6):387-396.**
76. **Hoff KJ, Lingner T, Meinicke P, Tech M: Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic acids research* 2009, 37(Web Server issue):W101-105.**
77. **Rho M, Tang H, Ye Y: FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research* 2010, 38(20):e191.**
78. **Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL: Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic acids research* 2012, 40(1):e9.**
79. **Eddy SR: A new generation of homology search tools based on probabilistic inference. *Genome informatics International Conference on Genome Informatics* 2009, 23(1):205-211.**
80. **Mulder N, Apweiler R: InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* 2007, 396:59-70.**
81. **Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J *et al*: The Pfam protein families database. *Nucleic acids research* 2012, 40(Database issue):D290-301.**
82. **Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al*: InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009, 37(Database issue):D211-215.**
83. **Richardson EJ, Escalettes F, Fotheringham I, Wallace RJ, Watson M: Meta4: a web application for sharing and annotating metagenomic gene predictions using web services. *Frontiers in genetics* 2013, 4:168.**
84. **Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M *et al*: IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic acids research* 2012, 40(Database issue):D123-129.**

85. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J *et al*: Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic acids research* 2011, 39(Database issue):D546-551.
86. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A *et al*: The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* 2008, 9:386.
87. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H: Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology* 2009, 4(4):265-270.
88. Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H: Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America* 2009, 106(19):7702-7707.
89. Stoddart D, Maglia G, Mikhailova E, Heron AJ, Bayley H: Multiple base-recognition sites in a biological nanopore: two heads are better than one. *Angewandte Chemie* 2010, 49(3):556-559.
90. Manrao EA, Derrington IM, Pavlenok M, Niederweis M, Gundlach JH: Nucleotide discrimination with DNA immobilized in the MspA nanopore. *PloS one* 2011, 6(10):e25723.
91. Watson M, Thomson M, Risse J, Talbot R, Santoyo-Lopez J, Gharbi K, Blaxter M: poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics* 2014.
92. Loman NJ, Quinlan AR: Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* 2014.
93. Lewis BP, Burge CB, Bartel DP: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005, 120(1):15-20.
94. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS: MicroRNA targets in Drosophila. *Genome biology* 2003, 5(1):R1.
95. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R: Fast and effective prediction of microRNA/target duplexes. *Rna* 2004, 10(10):1507-1517.
96. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M *et al*: Combinatorial microRNA target predictions. *Nature genetics* 2005, 37(5):495-500.
97. Wu X, Watson M: CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics* 2009, 25(6):832-833.
98. Creighton CJ, Nagaraja AK, Hanash SM, Matzuk MM, Gunaratne PH: A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *Rna* 2008, 14(11):2290-2296.
99. Mogilyansky E, Rigoutsos I: The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell death and differentiation* 2013, 20(12):1603-1614.
100. Gunaratne PH, Creighton CJ, Watson M, Tennakoon JB: Large-scale integration of MicroRNA and gene expression data for identification of enriched microRNA-mRNA associations in biological systems. *Methods Mol Biol* 2010, 667:297-315.
101. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S *et al*: The genome of a songbird. *Nature* 2010, 464(7289):757-762.

102. Gunaratne PH, Lin YC, Benham AL, Drnevich J, Coarfa C, Tennakoon JB, Creighton CJ, Kim JH, Milosavljevic A, Watson M *et al*: Song exposure regulates known and novel microRNAs in the zebra finch auditory forebrain. *BMC genomics* 2011, 12(1):277.

103. Vodovar N, Goic B, Blanc H, Saleh MC: In silico reconstruction of viral genomes from small RNAs improves virus-derived small interfering RNA profiling. *Journal of virology* 2011, 85(21):11016-11021.

104. Antoniewski C: Visitor, an informatic pipeline for analysis of viral siRNA sequencing datasets. *Methods Mol Biol* 2011, 721:123-142.

105. Watson M, Schnettler E, Kohl A: viRome: an R package for the visualization and analysis of viral small RNA sequence datasets. *Bioinformatics* 2013, 29(15):1902-1903.

106. Fonseca NA, Rung J, Brazma A, Marioni JC: Tools for mapping high-throughput sequencing data. *Bioinformatics* 2012, 28(24):3169-3177.

Appendix I – contribution to papers

First/Last author publications

WATSON, M. 2005. ProGenExpress: visualization of quantitative data on prokaryotic genomes. *BMC Bioinformatics,* 6**,** 98.
As sole author of this manuscript, I conceived the idea and carried out the research. I wrote and tested the software, I carried out all analyses, I interpreted the results, I wrote the paper and I responded to reviewers comments.
WATSON, M., DUKES, J., ABU-MEDIAN, A. B., KING, D. P. & BRITTON, P. 2007. DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. *Genome Biol,* 8**,** R190.
As first author of this manuscript, I conceived the idea and carried out the research. I wrote and tested the software, I carried out all analyses, I interpreted the results, I wrote the paper and I responded to reviewers comments. My co-authors assisted by providing data and by assisting with manuscript preparation.
WATSON, M., SCHNETTLER, E. & KOHL, A. 2013. viRome: an R package for the visualization and analysis of viral small RNA sequence datasets. *Bioinformatics*. 29, 1902-03
As first author of this manuscript, I conceived the idea and carried out the research. I wrote and tested the software, I carried out all analyses, I interpreted the results, I wrote the paper and I responded to reviewers comments. My co-authors assisted by providing data and by assisting with manuscript preparation. Please note that the journal Bioinformatics imposes a 2 page limit on software papers. However, this paper was subject to the full peer review process, including assessment by 3 independent peer reviewers. Additional research in this area is provided by Schnettler *et al* J Virol 2013, Schnettler *et al* J Gen Virl (2013), and Schnettler *et al* (submitted)
WU, X. & **WATSON, M**. 2009. CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics,* 25**,** 832-3.
As last author of this manuscript, I conceived the idea and helped carry out the research. I trained and supervised Dr Wu, I wrote an initial version of the software, I helped carry our and interpret data analyses, I helped write the paper and respond to reviews.

Please note that the journal Bioinformatics imposes a 2 page limit on software papers. However, this paper was subject to the full peer review process, including assessment by 2 independent peer reviewers. Additional research in this area is provided by Gunaratne *et al* Methods Mol Biol 2010, Warren *et al* Nature 2010, and Gunaratne *et al* BMC Genomics 2011
RICHARDSON, E. J., ESCALETTES, F., FOTHERINGHAM, I., WALLACE, R. J. & **WATSON, M**. 2013. Meta4: a web application for sharing and annotating metagenomic gene predictions using web services. *Front Genet,* 4**,** 168.
As last author of this manuscript, I conceived the idea and helped carry out the research. I trained and supervised Dr Richardson, I wrote an initial version of the software, I helped carry our and interpret data analyses, I helped write the paper and respond to reviews. Our co-authors provided data and advice, and helped with manuscript preparation.
YAO, Y., CHARLESWORTH, J., NAIR, V. & **WATSON, M**. 2013. MicroRNA expression profiles in avian haemopoietic cells. *Front Genet,* 4**,** 153.
As last author of this manuscript, I conceived the idea and helped carry out the research. I designed the experiment, carried out all data analyses, contributed to interpretation of the data and helped write the manuscript. Dr Yao carried out the wet-lab experiments (RNA extraction). Dr Charlesworth and Prof Nair assisted with interpretation and manuscript preparation.
YAO, Y., SMITH, L. P., NAIR, V. & **WATSON, M**. 2013. An avian retrovirus uses canonical expression and processing mechanisms to generate viral microRNA. *J Virol*. [Oct 2013 ahead of print]
As last author of this manuscript I jointly conceived the idea and helped carry out the research. This dataset was originally to be included in the Yao *et al* Front Genet 2013 paper; however, we discovered a very interesting microRNA which warranted separate publication. I carried out all data analyses, and therefore discovered the microRNA that the paper is based on. I helped with data interpretation and manuscript preparation.
RICHARDSON, E. J. & **WATSON**, M. 2013. The automatic annotation of bacterial genomes. *Brief Bioinform,* 14**,** 1-12.
This is a review. As last author, I conceived of the idea, and trained and supervised Dr Richardson. I

helped design the structure of the review, and guided Dr Richardson in writing the paper.
WATSON, M., THOMSON M., RISSE, J., TALBOT, R., SANTOYO-LOPEZ, J., GHARBI, K., BLAXTER, M. 2015. poRe: an R package for the visualization and analysis of nanopore sequencing data. Bioinformatics. 31(1):114-5
As first author of this manuscript, I conceived the idea and carried out the research. I wrote and tested the software, I carried out all analyses, I interpreted the results, I wrote the paper and I responded to reviewers comments. My co-authors assisted by providing data and by assisting with manuscript preparation.

Collaborative publications

YAO, Y., ZHAO, Y., XU, H., SMITH, L. P., LAWRIE, C. H., **WATSON, M**. & NAIR, V. 2008. MicroRNA profile of Marek's disease virus-transformed T-cell line MSB-1: predominance of virus-encoded microRNAs. *J Virol,* 82**,** 4007-15.
This paper forms part of a long-standing collaboration between my research group and Professor Nair's. As part of this project I carried out all data analysis, interpreted the data and helped write the manuscript.
YAO, Y., ZHAO, Y., SMITH, L. P., LAWRIE, C. H., SAUNDERS, N. J., **WATSON, M**. & NAIR, V. K. 2009. Differential expression of miRNAs in Marek's disease virus-transformed T-lymphoma cell lines. *J Gen Virol*., 90, 1551-59
This paper forms part of a long-standing collaboration between my research group and Professor Nair's. As part of this project I carried out all data analysis, interpreted the data and helped write the manuscript.
YAO, Y., ZHAO, Y., SMITH, L. P., **WATSON, M**. & NAIR, V. 2009. Novel microRNAs encoded by herpesvirus of turkeys (HVT): Evidence of miRNA evolution by duplication. *J Virol*., 83, 6969-73
This paper forms part of a long-standing collaboration between my research group and Professor Nair's. As part of this project I carried out all data analysis, interpreted the data and helped write the manuscript.
YAO, Y., SMITH, L. P., PETHERBRIDGE, L., **WATSON, M**. & NAIR, V. 2012. Novel microRNAs encoded by duck enteritis virus. *J Gen Virol,* 93**,** 1530-6.
This paper forms part of a long-standing collaboration between my research group and Professor

Nair's. As part of this project I carried out all data analysis, interpreted the data and helped write the manuscript.
SCHNETTLER, E., DONALD, C. L., HUMAN, S., **WATSON, M**., SIU, R. W., MCFARLANE, M., FAZAKERLEY, J. K., KOHL, A. & FRAGKOUDIS, R. 2013. Knockdown of piRNA pathway proteins results in enhanced Semliki Forest virus production in mosquito cells. *J Gen Virol*., 94, 1680-89
This paper forms part of a collaboration between my research group and Dr Kohl's. As part of this project I carried out all data analysis, interpreted the data and helped write the manuscript. Analyses were carried out using viRome (Watson *et al* 2013)
SCHNETTLER, E., RATINIER, M., **WATSON, M**., SHAW, A. E., MCFARLANE, M., VARELA, M., ELLIOTT, R. M., PALMARINI, M. & KOHL, A. 2013. RNA interference targets arbovirus replication in Culicoides cells. *J Virol,* 87**,** 2441-54.
This paper forms part of a collaboration between my research group and Dr Kohl's. As part of this project I carried out all data analysis, interpreted the data and helped write the manuscript. Analyses were carried out using viRome (Watson *et al* 2013)
SCHNETTLER, E., TYKALOVA, H., **WATSON, M**., SHARMA, M., STERKEN M. G., OBBARD, D. J., LEWIS, S. H., MCFARLANE, M., BELL-SAKYI, L., BARRY, G., WEISHEIT, S., BEST, S. M., KUHN, R. J., PIJLMAN, G. P., CHASE-TOPPING, M. E., GOULD, E. A., GRUBHOFFER, L., FAZAKERLEY, J. K., KOHL, A. (2014) Induction and suppression of tick cell antiviral RNAi responses by tick-borne flaviviruses, *Nucleic Acids Research* 42(14):9436-46
This paper forms part of a collaboration between my research group and Dr Kohl's. As part of this project I carried out all data analysis, interpreted the data and helped write the manuscript. Analyses were carried out using viRome (Watson *et al* 2013)
GUNARATNE, P. H., CREIGHTON, C. J., **WATSON, M**. & TENNAKOON, J. B. 2010. Large-scale integration of MicroRNA and gene expression data for identification of enriched microRNA-mRNA associations in biological systems. *Methods Mol Biol,* 667**,** 297-315.
This paper is a review written by all authors that explores methods for the microRNA-mRNA data integration in more depth. CORNA (Wu and Watson, 2009) is featured in this manuscript.
WARREN, W. C., CLAYTON, D. F., ELLEGREN, H., ARNOLD, A. P., HILLIER, L. W., KUNSTNER, A., SEARLE,

S., WHITE, S., VILELLA, A. J., FAIRLEY, S., HEGER, A., KONG, L., PONTING, C. P., JARVIS, E. D., MELLO, C. V., MINX, P., LOVELL, P., VELHO, T. A., FERRIS, M., BALAKRISHNAN, C. N., SINHA, S., BLATTI, C., LONDON, S. E., LI, Y., LIN, Y. C., GEORGE, J., SWEEDLER, J., SOUTHEY, B., GUNARATNE, P., **WATSON, M**., NAM, K., BACKSTROM, N., SMEDS, L., NABHOLZ, B., ITOH, Y., WHITNEY, O., PFENNING, A. R., HOWARD, J., VOLKER, M., SKINNER, B. M., GRIFFIN, D. K., YE, L., MCLAREN, W. M., FLICEK, P., QUESADA, V., VELASCO, G., LOPEZ-OTIN, C., PUENTE, X. S., OLENDER, T., LANCET, D., SMIT, A. F., HUBLEY, R., KONKEL, M. K., WALKER, J. A., BATZER, M. A., GU, W., POLLOCK, D. D., CHEN, L., CHENG, Z., EICHLER, E. E., STAPLEY, J., SLATE, J., EKBLOM, R., BIRKHEAD, T., BURKE, T., BURT, D., SCHARFF, C., ADAM, I., RICHARD, H., SULTAN, M., SOLDATOV, A., LEHRACH, H., EDWARDS, S. V., YANG, S. P., LI, X., GRAVES, T., FULTON, L., NELSON, J., CHINWALLA, A., HOU, S., MARDIS, E. R. & WILSON, R. K. 2010. The genome of a songbird. *Nature,* 464**,** 757-62.
This paper is the culmination of a long-term, international project to sequence the genome of the zebrafinch. I provided key data analyses and software tools that provided information on the function of key datasets. I analysed data and provided tools so that others could analyse data. Note that CORNA (Wu and Watson, 2009) was used extensively, and is featured/cited in the supplementary methods to the paper.
GUNARATNE, P. H., LIN, Y. C., BENHAM, A. L., DRNEVICH, J., COARFA, C., TENNAKOON, J. B., CREIGHTON, C. J., KIM, J. H., MILOSAVLJEVIC, A., **WATSON, M**., GRIFFITHS-JONES, S. & CLAYTON, D. F. 2011. Song exposure regulates known and novel microRNAs in the zebra finch auditory forebrain. *BMC Genomics,* 12**,** 277.
I carried out data analysis and interpretation, as well as contributing to manuscript preparation. Data analyses were carried out using CORNA (Wu and Watson, 2009)
HARVEY, P. C., **WATSON, M**., HULME, S., JONES, M. A., LOVELL, M., BERCHIERI, A., JNR., YOUNG, J., BUMSTEAD, N. & BARROW, P. 2011. Salmonella enterica serovar Typhimurium colonizing the lumen of the chicken intestine are growing slowly and up-regulate a unique set of virulence and metabolism genes. *Infect. Immun.***,** IAI.01390-10.
I carried out all data analyses, helped with interpretation and contributed to the manuscript.
RICHARDSON, E. J., LIMAYE, B., INAMDAR, H., DATTA, A., MANJARI, K. S., PULLINGER, G. D., THOMSON, N. R., JOSHI, R. R., **WATSON, M**. & STEVENS, M. P. 2011. Genome Sequences of Salmonella enterica Serovar Typhimurium, Choleraesuis, Dublin, and Gallinarum Strains of Well-Defined Virulence in Food-Producing Animals. *J Bacteriol,* 193**,** 3162-3.
I conceived the study and helped carry out the research. I supervised Dr Richardson (first author) with Prof. Stevens (last author). Whilst this is a genome announcement (limited to 500 words), it is

reviewed by the editor of the journal. Furthermore, this paper represents the culmination of over 12 months of work in which we sequenced and annotated 4 *Salmonella* genomes, strains of well-defined virulence in farm animals. Dr's Limaye, Inamdar, Datta, Manjari and Joshi joined the project as part of a BBSRC India Partnering Award on which I was PI from 2008-2011
CLAYTON, D. J., BOWEN, A. J., HULME, S. D., BUCKLEY, A. M., DEACON, V. L., THOMSON, N. R., BARROW, P. A., MORGAN, E., JONES, M. A., **WATSON, M**. & STEVENS, M. P. 2008. Analysis of the role of 13 major fimbrial subunits in colonisation of the chicken intestines by Salmonella enterica serovar Enteritidis reveals a role for a novel locus. *BMC Microbiol,* 8**,** 228.
I co-supervised Dr Clayton (first author) with Prof Mark Stevens (last author). As such I provided training and supervision to Dr Clayton and oversaw all data analysis and interpretation that she carried out as part of the project. I helped write the manuscript and responded to reviewers
THOMSON, N. R., CLAYTON, D. J., WINDHORST, D., VERNIKOS, G., DAVIDSON, S., CHURCHER, C., QUAIL, M. A., STEVENS, M., JONES, M. A., **WATSON, M**., BARRON, A., LAYTON, A., PICKARD, D., KINGSLEY, R. A., BIGNELL, A., CLARK, L., HARRIS, B., ORMOND, D., ABDELLAH, Z., BROOKS, K., CHEREVACH, I., CHILLINGWORTH, T., WOODWARD, J., NORBERCZAK, H., LORD, A., ARROWSMITH, C., JAGELS, K., MOULE, S., MUNGALL, K., SANDERS, M., WHITEHEAD, S., CHABALGOITY, J. A., MASKELL, D., HUMPHREY, T., ROBERTS, M., BARROW, P. A., DOUGAN, G. & PARKHILL, J. 2008. Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res,* 18**,** 1624-37.
I co-supervised Dr Clayton (second author) with Prof Paul Barrow (3rd last author). As such I provided training and supervision to Dr Clayton and oversaw all data analysis and interpretation that she carried out as part of the project.

Appendix II – full papers

JOURNAL OF VIROLOGY, Apr. 2008, p. 4007–4015 Vol. 82, No. 8
0022-538X/08/$08.00+0 doi:10.1128/JVI.02659-07

MicroRNA Profile of Marek's Disease Virus-Transformed T-Cell Line MSB-1: Predominance of Virus-Encoded MicroRNAs▽

Yongxiu Yao,[1] Yuguang Zhao,[1] Hongtao Xu,[1] Lorraine P. Smith,[1] Charles H. Lawrie,[2] Michael Watson,[1] and Venugopal Nair[1]*

Division of Microbiology, Institute for Animal Health, Compton, Berkshire RG20 7NN, United Kingdom,[1] and LRF Molecular Haematology Unit, Nuffield Department of Clinical Laboratory Sciences, University of Oxford, Oxford OX3 9DU, United Kingdom[2]

Received 14 December 2007/Accepted 25 January 2008

Research over the last few years has demonstrated the increasing role of microRNAs (miRNAs) as major regulators of gene expression in diverse cellular processes and diseases. Several viruses, particularly herpesviruses, also use the miRNA pathway of gene regulation by encoding their own miRNAs. Marek's disease (MD) is a widespread lymphomatous neoplastic disease of poultry caused by the highly contagious Marek's disease virus type 1 (MDV-1). Recent studies using virus-infected chicken embryo fibroblasts have identified at least eight miRNAs that map to the R_L/R_S region of the MDV genome. Since MDV is a lymphotropic virus that induces T-cell lymphomas, analysis of the miRNA profile in T-cell lymphoma would be more relevant for examining their role in oncogenesis. We determined the viral and host miRNAs expressed in MSB-1, a lymphoblastoid cell line established from an MDV-induced lymphoma of the spleen. In this paper, we report the identification of 13 MDV-1-encoded miRNAs (12 by direct cloning and 1 by Northern blotting) from MSB-1 cells. These miRNAs, five of which are novel MDV-1 miRNAs, map to the Meq and latency-associated transcript regions of the MDV genome. Furthermore, we show that miRNAs encoded by MDV-1 and the coinfected MDV-2 accounted for >60% of the 5,099 sequences of the MSB-1 "miRNAome." Several chicken miRNAs, some of which are known to be associated with cancer, were also cloned from MSB-1 cells. High levels of expression of MDV-1-encoded miRNAs and potentially oncogenic host miRNAs suggest that miRNAs may have major roles in MDV pathogenesis and neoplastic transformation.

Marek's disease (MD) is a naturally occurring rapid-onset aggressive T-cell lymphoma of poultry. Named after the Hungarian veterinarian Jószef Marek, who first reported the disease in 1907 (41), the disease is induced by Marek's disease virus type 1 (MDV-1), a highly contagious alphaherpesvirus belonging to the genus *Mardivirus* of the family *Herpesviridae* (31). Apart from being a major disease affecting poultry health and welfare, MD is considered to be an excellent biomedical model for virus-induced lymphoma (7, 14). Among the 100-plus genes predicted for the MDV genome (40, 47, 48), the gene for the basic leucine zipper protein Meq is considered to be the major oncogene (39, 44). Some of the functions of Meq associated with oncogenic properties, such as its interaction with CtBP, parallel those of other viral oncogenic sequences, such as adenovirus E1A and Epstein-Barr virus (EBV) nuclear antigens EBNA3A and -3C (6), highlighting the convergent evolution of oncogenic pathways in these viruses. Recent studies have also identified the role of other genes, such as the pp38 (23), viral interleukin-8 (vIL-8) (49), ICP4 (15, 38), R-LORF4 (33), UL36 (32), and MDV-encoded telomerase RNA (22, 63) genes, in pathogenesis.

Increasing evidence demonstrates that in addition to the direct role of protein-encoding genes, noncoding RNAs have profound effects in mediating neoplastic transformation (13). Among these, the 22-nucleotide microRNAs (miRNAs) have emerged as a major regulatory tier of gene expression, with the potential of targeting up to 30% of genes in humans (17, 27, 37). Given their small size with the capability for regulating multiple genes, several viruses have adopted the miRNA machinery to manipulate the cellular and viral pathways of gene regulation by encoding their own miRNAs (19, 24, 26, 42). Among the different families of viruses, herpesviruses have exploited the miRNA-mediated gene regulation pathway most successfully, since 124 of the 127 virus-encoded miRNAs in miRBase release 10.1 (http://microrna.sanger.ac.uk) are encoded by herpesviruses. It has been suggested that the miRNA-mediated regulatory mechanisms are very suited for the herpesvirus life cycle, which is characterized by nuclear replication and latent periods with minimal antigen expression (19).

Specific miRNA signatures in different types of tumors have been identified using high-throughput microarray analysis of miRNA expression (60, 64, 67). Considering the aggressive nature and rapid onset of tumors induced by MDV-1, analysis of the miRNA profile of MDV-transformed tumor cells could provide further insights into MD oncogenesis. Previous studies using small RNAs from MDV-infected chicken embryo fibroblasts (CEF) have identified several miRNAs, including eight MDV-encoded miRNAs that mapped to the Meq and the latency-associated transcript (LAT) region of the genome (8, 9). Although identification of MDV and host miRNAs in lytically infected CEF is valuable, understanding the expression profiles of miRNAs in the lymphocyte target cells of MD lymphomas would be crucial to delineate their role in neoplas-

* Corresponding author. Mailing address: Division of Microbiology, Institute for Animal Health, Compton, Berkshire, United Kingdom RG20 7NN. Phone: 441635 577356. Fax: 441635 577263. E-mail: venu.gopal@bbsrc.ac.uk.

▽ Published ahead of print on 6 February 2008.

TABLE 1. Sequences, chromosomal locations, and cloning frequencies of chicken miRNAs cloned from an MSB-1 library

Name[a]	Sequence	No. of hits in library	Chromosomal location	Start position	End position
gga-mir-7	TGGAAGACTAGTGATTTTGTTG	3	Z_random	12717978	12718086
gga-mir-15a	TAGCAGCACATAATGGTTTGT	28	1	161540787	161540869
gga-mir-15b	TAGCAGCACATCATGGTTTGCA	9	9	21649291	21649381
			1	161540645	161540728
gga-mir-16	TAGCAGCACGTAAATATTGGTG	82	9	21649116	21649209
gga-mir-17-5p	CAAAGTGCTTACAGTGCAGGTAA	55	1	140631124	140631208
gga-mir-18a	TAAGGTGCATCTAGTGCAGATA	10	1	140630969	140631061
gga-mir-18b	TAAGGTGCATCTAGTGCAGTTA	10	4	3781954	3782037
gga-mir-19a	TGTGCAAATCTATGCAAAACTGA	71	1	140630835	140630915
gga-mir-19b	TGTGCAAATCCATGCAAAACTGA	71	1	140630526	140630612
gga-mir-20a	TAAAGTGCTTATAGTGCAGGTAG	18	1	140630649	140630746
gga-mir-20b	CAAAGTGCTCATAGTGCAGGTAG	16	4	3781773	3781857
gga-mir-21	TAGCTTATCAGACTGATGTTGA	249	19	6933581	6933677
gga-mir-23b	ATCACATTGCCAGGGATTAC	2	Z_random	14203199	14203284
gga-mir-24	TGGCTCAGTTCAGCAGGAACAG	6	Z_random	14203968	14204035
gga-mir-26a	TTCAAGTAATCCAGGATAGGC	14	2	4034213	4034289
gga-mir-27b	TTCACAGTGGCTAAGTTCTGC	10	Z_random	14203435	14203531
			26	1280249	1280328
gga-mir-29b	TAGCACCATTTGAAATCAGTGTT	98	1	204569	204649
gga-mir-30b	TGTAAACATCCTACACTCAGCT	1	2	141145952	141146038
			3	80699454	80699525
gga-mir-30c	TGTAAACATCCTACACTCTCAGCT	1	23	4431115	4431203
gga-mir-30d	TGTAAACATCCCCGACTGGAAGC	16	2	141142201	141142264
gga-mir-30a-5p	TGTAAACATCCTCGACTGGAAGCT	13	3	80674840	80674911
gga-mir-33	GTGCATTGTAGTTGCATTG	6	1	46203134	46203202
gga-mir-34a	TGGCAGTGTCTTAGCTGGTTGTT	7	21	3118925	3119033
gga-miR-92	TATTGCACTTGTCCCGGCCTGT	9	1	140630413	140630490
gga-mir-101	TACAGTACTGTGATAACTGAAG	9	Z	11651019	11651097
gga-mir-106	AAAAGTGCTTACAGTGCAGGTA	55	4	3782085	3782165
gga-mir-130a	CAGTGCAATATTAAAAGGGCA	4	15	393029	393111
gga-mir-140	AGTGGTTTTACCCTATGGTAG	5	11_random	250924	251018
gga-mir-142-3p	TGTAGTGTTTCCTACTTTATGGA	224	Un	130069949	130070036
gga-mir-142-5p	CCCATAAAGTAGAAAGCACTAC	243	Un	130069949	130070036
gga-mir-146b	TGAGAACTGAATTCCATAGGCG	44	6	22212586	22212690
gga-mir-181a	AACATTCAACGCTGTCGGTGAGTT	5	8	1957561	1957664
			17	945791	945881
			8	1957750	1957838
gga-mir-181b	AACATTCATTGCTGTCGGTGGGTTT	6	17	944157	944241
gga-mir-221	AGCTACATTGTCTGCTGGGTTTC	11	1	104369325	104369423
gga-mir-222	AGCTACATCTGGCTACTGGGT	7	1	104368821	104368918
gga-mir-301	CAGTGCAATAATATTGTCAAAGCATT	3	15	391803	391895
gga-mir-456	CAGGCTGGTTAGATGGTTGTCCT	4	3	27735429	27735540
gga-let-7i	TGAGGTAGTAGTTTGTGCTGT	148	1	29988296	29988379
			12	6149732	6149821
gga-let-7a	TGAGGTAGTAGGTTGTATAGTT	12	24	3263054	3263125
			1	67795667	67795742
gga-mir-363*	AATTGCACGGTATCCATCTGTA	30			
gga-mir-454*	TAGTGCAATATTGCTTATAGGGT	5			
gga-mir-425*	AATGACACGATCACTCCCGTTGA	6			
gga-mir-191*	CAACGGAATCCCAAAAGCAGCTG	8			
gga-mir-22*	AAGCTGCCAGTTGAAGAACTGT	6			
gga-mir-739*	AAGGCCGAAGTGGAGAAGGGTTCCA	1			

[a] *, novel miRNAs identified in chickens and assigned names on the basis of homology to miRNAs in other species.

tic transformation. Primary MD lymphomas are often heterogeneous mixtures of neoplastic T cells and nontransformed cells of other lineages (50), so analysis of the whole tumor may not provide the miRNA profile of the transformed target cell. However, the ability to establish homogeneous clonal populations of lymphoblastoid cell lines from primary tumors has helped to gain insights into the gene expression profiles of MD tumor cells (10). We reasoned that examination of the miRNA profiles of MDV-transformed lymphoblastoid cell lines could help to analyze their roles in neoplastic transformation and in the maintenance of MDV latency in target T cells.

MSB-1 is an MDV-transformed $CD4^+$ T-cell line derived from a spleen lymphoma induced by the BC-1 strain of MDV-1 (1, 30). The MSB-1 cell line, used in this study at passage level 13, has a $CD4^+$ phenotype and has been shown to be coinfected with MDV-1 and MDV-2 (66). The MSB-1 cell line has both integrated and circular copies of the MDV-1 genome (56) and induces tumors when it is inoculated into susceptible

TABLE 2. Sequences and genomic positions of MDV-1 miRNAs[a]

Name	Sequence (5′ to 3′)	Length (nt)	No. of hits	Nucleotide position[b]
MDV1-miR-M1-5p	UGCUUGUUCACUGUGCGGCA(UUAU)	20–24	339	136873–136896
MDV1-miR-M1-3p	(A)UGCUGCGCAUGAAAGAGCGA(A)	21–23	4	136913–136934
MDV1-miR-M2-5p	(G)UUGUAUUCUGCCCGGUAGUCC(GUUU)	22–26	16	134231–134256
MDV1-miR-M2-3p	(A)CGGACUGCCGCAGAAUAGC(UUU)	19–22	11	134270–134292
MDV1-miR-M3-5p	(CAUG)AAAAUGUGAAACCUCUC(CCGCU)	20–25	390	134079–134104
MDV1-miR-M4-5p	(UUUAA)UGCUGUAUCGGAACC(CUUCGUU)	19–26	341	134367–134393
MDV1-miR-M4-3p	(CGA)AUGGUUCUGACAGCAUGAC(CU)	20–22	50	134403–134426
MDV1-miR-M5-5p	AACCGUAUGCGAUCACAUUGAC	22	0	133606–133628
MDV1-miR-M5-3p	(U)GUGUAUCGUGGUCGUCUACU(GUU)	21–24	176	133647–133670
MDV1-miR-M6-5p	(UCU)GUUGUUCCGUAGUGUUC(UC)	18–22	278	142335–142356
MDV1-miR-M6-3p	(GAG)AUCCCUGCGAAAUGACAGU(U)	19–23	14	142370–142392
MDV1-miR-M7-5p	(UGU)UAUCUCGGGGAGAUCCC(GAU)	19–23	800	142508–142530
MDV1-miR-M8-5p	UAUUGUUCUGUGGUUGGUUUC(GA)	21–23	18	142216–142238
MDV1-miR-M8-3p	(GU)GACCUCUACGGAACAAUAG(U)	20–22	36	142258–142279
MDV1-miR-M9-5p	UUUUCUCCUUCCCCCCGGAGUU(CA)	22–24	45	133374–133397
MDV1-miR-M9-3p	AAACUCCGAGGGCAGGAAAAAG	22	1	133414–133435
MDV1-miR-M10-3p	(UCG)AAAUCUCUACGAGAUAACA(GU)	20–23	6	142667–142690
MDV1-miR-M11-5p	UUUUCCUUACCGUGUAGCUUAGA	23	2	136053–136075
MDV1-miR-M11-3p	UGAGUUACAUGGUCAGGGGAUU	22	0	136092–136113
MDV1-miR-M12-3p	(U)UGCAUAAUACGGAGGGUUCU(G)	21–22	35	133925–133946
MDV1-miR-13-3p	GCAUGGAAACGUCCUGGGAAA	21	0	142313–142333

[a] Sequence variation surrounding the recovered BC-1 miRNAs is indicated by parentheses surrounding the variable nucleotides. miRNAs derived from a single primary miRNA stem-loop precursor are indicated by a "-5p" (5′ arm) or "-3p" (3′ arm) suffix.
[b] Based on the Md5 sequence (GenBank accession no. AF243438).

chickens (21, 35). As reported for some MD tumors, these cells also showed truncated forms of p53 tumor suppressor protein (62). Northern blot analysis of the expression of MDV-induced miRNAs in MSB-1 cells showed that many of the miRNAs are expressed at much higher levels than those in infected CEF (8, 9). These results demonstrate that the MSB-1 lymphoblastoid cell line, which shares many properties of MD tumors, could be used as a model system for analyzing the molecular pathways and mechanisms of neoplastic transformation in MD tumors. We recently reported the construction of a library, using small RNAs fractionated from MSB-1 cells, to identify novel MDV-2-encoded miRNAs (66). In this paper, we describe the results of analysis of the MSB-1 "miRNAome" to examine the population of host and viral miRNAs expressed in this transformed cell line.

MATERIALS AND METHODS

Cells and viruses. CEF prepared from 10-day-old specific-pathogen-free embryos obtained from flocks maintained at the Institute for Animal Health were used for the propagation of viruses. Low-passage-number virus stocks of RB-1B (58) grown in CEF for 72 to 96 h were used for the preparation of RNA for Northern blotting analysis. The MDV-transformed lymphoblastoid cell line MSB-1 (1) and the REV-T-transformed (16) chicken $CD4^+$ T-cell line AVOL-1 were grown at 38.5°C in 5% CO_2 in RPMI 1640 medium containing 10% fetal calf serum, 10% tryptose phosphate broth, and 1% sodium pyruvate.

Cloning and identification of miRNAs. We have previously described the construction of a cDNA library from small RNAs prepared from MSB-1 cells (66). Concatemerized sequences of putative miRNAs from the pGEM-T Easy (Promega, Southampton, United Kingdom) library were determined using vector-specific primers. High-quality reads of small RNA sequences with both 5′ and 3′ adapters were analyzed for the characterization of miRNAs.

Northern blotting analysis. Total RNA was extracted from cultured cells by using TRIzol reagent (Invitrogen) according to standard methods described by the manufacturer. RNAs were also isolated from samples of MD lymphomas as well as from livers, brains, hearts, kidneys, ovaries, lungs, thymuses, and spleens of uninfected adult chickens, using TRIzol reagent. Samples of 20 μg total RNA were resolved in a 15% polyacrylamide-urea gel and blotted onto a GeneScreen Plus membrane (Perkin-Elmer). DNA oligonucleotides with the exact complementary sequence to selected miRNAs were end labeled with [γ-^{32}P]ATP by use of T4 polynucleotide kinase (New England Biolabs, Hertfordshire, United Kingdom) to generate high-specific-activity probes. Hybridization, washing, and autoradiography were carried out as previously described (36, 53).

RESULTS

Identification of miRNAs expressed in MSB-1 cells. The MDV-transformed lymphoblastoid T-cell line MSB-1 has been used extensively in different laboratories for various studies, particularly for the analysis of MDV latency programs (43). As a tumor cell line latently infected with MDV-1, we chose MSB-1 to analyze the miRNA profile of MD tumor cells. Sequence analysis of ~1,200 pGEM-T Easy clones of cDNA concatemers of small RNA sequences from the MSB-1 library identified a total of 5,099 high-quality reads. The sequences were scored as miRNAs on the basis that their flanking sequences could be predicted into a stem-loop structure with low free energyBlast (2). Homology searches of these sequences against the miRBase (25) and GenBank (5) databases were used to determine the identities of the different host- and virus-encoded small RNAs. Of the total reads, 1,641 (32.2%) matched known *Gallus gallus* miRNAs in the miRBase. The most abundant host miRNAs in the MSB-1 library were gga-miR-21, gga-miR-142-3p, gga-miR-142-5p, gga-let-7i, gga-miR-29b, gga-miR-16 gga-miR-17-5p, gga-miR-19a, gga-miR-19b, gga-miR-106, and gga-miR-146b. Sequence analysis of the clones from the MSB-1 library also identified six novel chicken miRNAs, which appeared to be homologs of hsa-miR-363, hsa-miR-454-3p, hsa-miR-425-5p, bta-miR-191, hsa-miR-22, and dre-miR-739. The number of reads of each host-encoded miRNA in the library and their chromosomal locations are shown in Table 1. In addition to these miRNAs and the virus-encoded miRNAs (see below), 128 (2.5%) were noncoding

RNA fragments, 76 (1.5%) were mRNA fragments, and 174 (3.4%) showed no matches to any known RNAs.

For the identification of the virus-encoded miRNAs, BLAST searches were carried out against the full-length sequences of MDV-1 (Md5 strain; GenBank accession number AF243438) and MDV-2 (HPRS-24 strain; GenBank accession number AB049735). We have previously shown that 518 (10.2%) sequences from the MSB-1 library are encoded by MDV-2 (66). However, the majority of the 2,562 (50.2%) sequences from this library showed perfect sequence identity to the genome sequence of the Md5 strain. These miRNA sequences, ranging in length from 18 to 26 nucleotides, belonged to 12 distinct MDV-1-encoded miRNAs. These included the eight miRNAs (mdv-miR-M1 to -M8) identified previously from MDV-infected CEF cultures (8) and four novel MDV-1-encoded miRNAs (mdv-miR-M9 to -M12). Additionally, we identified another novel miRNA, mdv-miR-M13, using Northern blotting analysis of RNAs extracted from MSB-1 cells (see below). The genomic location and cloning frequency of each of the MDV-1-encoded miRNAs are shown in Table 2.

MDV-1 miRNAs fold into distinct hairpin structures. For the validation of a sequence as a miRNA, demonstration of its expression as well as its processing through the miRNA biogenesis pathway is required. One of the distinct indicators of miRNA biogenesis is the presence of adjacent complementary sequences that are able to form stable hairpins. In order to analyze the potential precursor structures of miRNAs encoded by MDV-1, the sequences of the 13 miRNAs with their adjacent 60 to 80 nucleotides were analyzed by MFOLD calculation, and the secondary structures were drawn using RNADRAW software as described previously (66). All of the MDV-1 miRNAs showed a stable hairpin with long paired stems (Fig. 1), indicating that they are bona fide miRNAs. Of the two strands of the miRNA duplex generated during biogenesis, only one of the strands, the miRNA strand, is incorporated into the RNA-induced silencing complex and guides gene regulation (3). Although the non-miRNA strand is rapidly degraded, in many instances it is also captured during cloning and may sometimes be detected with a comparable frequency to that of the miRNA strand (66). Among the MDV-1 miRNAs in MSB-1 cells, two mature forms, representing both strands of the duplex, were demonstrated by cloning or Northern blotting for 8 of the 13 candidate miRNAs, increasing the total number of miRNAs to 21. The suffixes "-5p" and "-3p" were added to designations to indicate the 5′ and 3′ arms, respectively, of the stem-loop precursor from which the miRNAs were derived (Table 2).

MDV-1 miRNAs show differences in cloning frequencies. We then examined the cloning frequency of each of the MDV-1 miRNAs as a measure of their expression levels in MSB-1 cells. The most abundantly cloned miRNAs were mdv-miR-M7-5p (800 hits), mdv-miR-M3-5p (390 hits), mdv-miR-M4-5p (341 hits), mdv-miR-M1-5p (339 hits), mdv-miR-M6-5p (278 hits), and mdv-miR-M5-3p (176 hits). Compared to this, mdv-miR-M10, -M11, and -M13 were of very low abundance, while mdv-miR-M2, -M8, -M9, and -M12 showed moderate copy numbers in the library. For most miRNAs, the non-miRNA strand of the duplex was either not cloned or had a relative frequency much lower than that of the miRNA strand. However, for some of the miRNAs, such as mdv-miR-M2, the

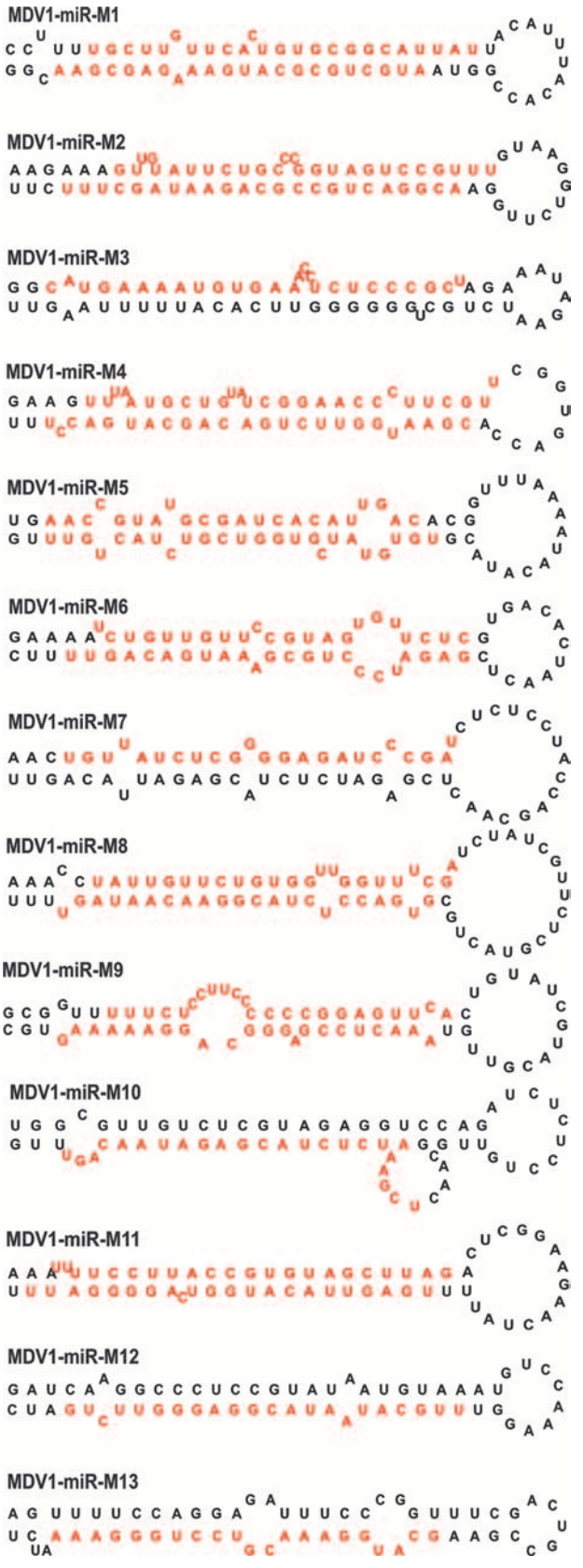


FIG. 1. Secondary structures of MDV-1 pre-miRNAs predicted using the MFOLD algorithm (68). The mature miRNA strands are indicated in red.

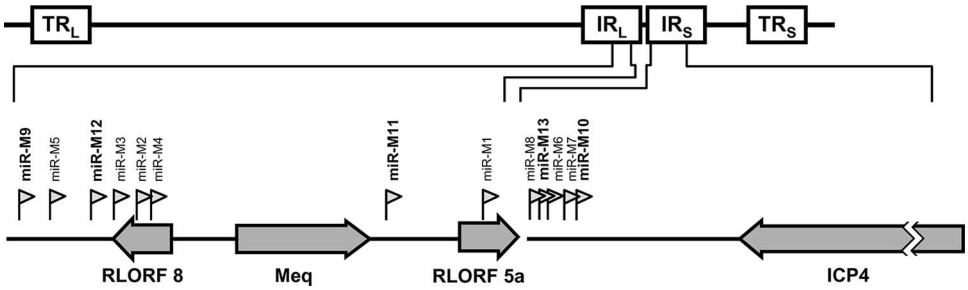


FIG. 2. Genomic locations of MDV-1 miRNAs. The schematic diagram shows where the MDV-1 miRNAs (small arrowheads) identified in this report map. The five novel miRNAs identified in this report are shown in bold. The TR_L and IR_L regions flanking the unique long region and the TR_S and IR_S regions flanking the unique short regions are shown. Genomic positions and orientations of MDV ORFs contained in the miRNA loci are indicated.

frequencies of both strands were very similar, suggesting that both may be functional and are incorporated into the RNA-induced silencing complex.

MDV-1 miRNAs are clustered in the repeat regions of the MDV-1 genome. The nucleotide sequence positions of the different miRNAs are shown in Table 2. All 13 MDV-1-encoded miRNAs are clustered in an ~9-kb region within the R_L/R_S sequences of the MDV genome (Fig. 2). The miRNAs mdv-miR-M2, -M3, -M4, -M5, -M9, and -M12 are located upstream of Meq and are antisense to the R-LORF8 transcript. The miRNAs mdv-miR-M1 and -M11 lie downstream of Meq and are embedded within the open reading frame (ORF) of the L1/LORF5a transcript (46, 59) as well as within the intron of the splice variant Meq-sp (51). MDV-1-encoded miRNAs mdv-miR-M6, -M7, -M8, -M10, and -M13 are located between the a-like sequence and the ICP4 sequence within the large intron of the LAT of MSR (15). Of this cluster, the miRNAs mdv-miR-M6 and mdv-miR-M13 were separated by only a single nucleotide. The occurrence of the miRNAs in distinct clusters in the same orientation strongly suggests that these miRNAs are likely to be processed as multicistronic pre-miRNA transcripts. Despite being processed from a single transcript, there are differences in the expression levels of the mature miRNAs, and these are thought to be due to differences in Drosha processing and/or miRNA stability.

Analysis of miRNA expression by Northern blotting. For further confirmation of the expression of miRNAs in MSB-1 cells, Northern blot hybridization with individual MDV-1 miRNA probes was carried out on RNAs extracted from MSB-1, AVOL-1 (an MDV-negative T-cell line), or uninfected or RB-1B virus-infected CEF cells and from samples of MD lymphoma. These studies confirmed that MDV-1-encoded miRNAs are expressed at high levels in MSB-1 cells and MD lymphomas and at low levels in infected CEF (Fig. 3). No signals were obtained from the RNAs extracted from AVOL-1 cells and uninfected CEF, validating the specificity of the miRNA probes. Based on the intensities of signals, the levels of expression of the majority of miRNAs were similar in both MSB-1 cells and lymphoma samples. Some of the most abundantly cloned miRNAs, such as mdv-miR-M3, -M4, -M5, -M6, and -M7, showed very strong signals by Northern blotting, validating the correlation between cloning frequency and expression level. Similarly, mdv-miR-M10, -M11, and -M13, cloned at very low frequencies from the library, gave weak signals by Northern blotting. A previous study using Northern blotting of RNAs extracted from MD lymphomas reported that mdv-miR-M6 is expressed at low levels, and mdv-miR-M7 was not detected at all (8). However, our studies using probes specific for the mdv-miR-M6-5p and mdv-miR-M7-5p strands gave strong signals for all samples, including MD lymphomas, indicating that the -5p strand of the duplex is the functional miRNA strand. The failure to detect these miRNAs in Northern blots in the previous study was likely due to the use of the non-miRNA strand as the probe. In most cases, both pre-miRNAs and mature miRNAs could be detected by Northern blotting, with the former giving much lower signals. For some of the miRNAs, such as mdv-miR-M5-5p, mdv-miR-M9-3p, and mdv-miR-M13, the signals of pre-miRNAs were higher than those of the mature miRNAs, indicating less efficient processing.

Northern blotting was also carried out on RNAs extracted from eight different normal tissues from adult noninfected chickens to validate the expression of some of the miRNAs cloned from the MSB-1 library. Some of these miRNAs, including novel chicken miRNAs such as gga-miR-363, gga-miR-454 gga-miR-425, gga-miR-191, and gga-miR-22, could be detected by Northern blot analysis, albeit with differences in expression levels between tissues (Fig. 3b). While gga-miR-425 and gga-miR-22 showed high levels of expression in all tissues, gga-miR-454 was detected at very low levels. The expression of the miRNAs gga-miR-191, gga-miR-363, and gga-miR-425 in lymphoid organs, namely, the spleen, thymus, and the lungs (55), was at the levels observed for the lymphocyte-specific miRNA gga-miR-142 (54, 65), which gave strong signals with probes specific for either of the strands of the duplex in these tissues.

DISCUSSION

As efficient inducers of cancer, oncogenic viruses have helped to reveal several major pathways of oncogenesis. Most of these pathways involve the interactions of virus-encoded oncoproteins, such as simian virus 40 T antigen, adenovirus E1A, human papillomavirus E6/E7, and EBV EBNAs (69). In MD tumors, MDV-encoded Meq is considered to be the major oncoprotein (39), although other proteins also contribute to oncogenesis (44). The discovery of virus-encoded miRNAs in several oncogenic viruses (53) has added yet another armory to

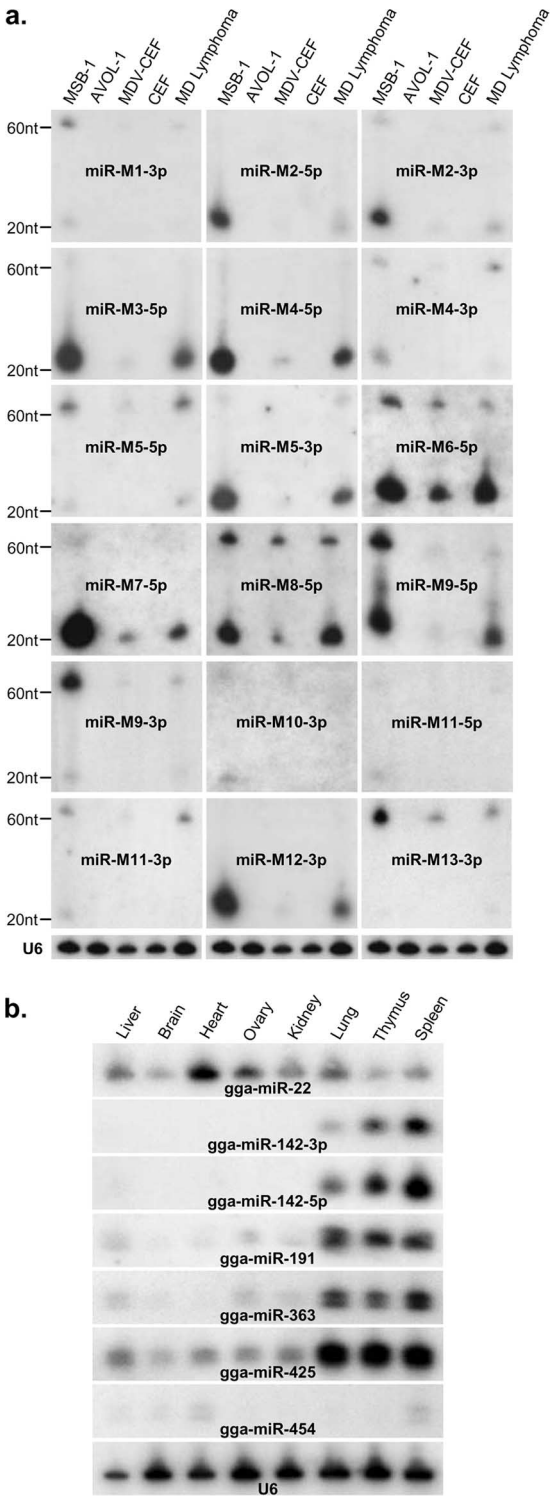


FIG. 3. (a) Northern blotting analysis for determining the expression of MDV-1 miRNAs. Twenty micrograms of total RNA from MSB-1 cells, the MDV-negative lymphoid cell line AVOL-1, MDV-1-infected CEF, uninfected CEF, or MD lymphoma tissues was separated in a 15% denaturing polyacrylamide gel, blotted, and probed with end-labeled antisense oligonucleotides to the indicated miRNAs. Size markers to indicate the positions of the pre-miRNA and the mature miRNA are shown. The cellular U6 snRNA served as the loading control, and a representative blot of this set is shown. (b) Analysis of tissue-specific expression of *Gallus gallus* (gga-) miRNAs identified in the MSB-1 library. Total RNAs (20 μg) extracted from different tissues of chickens were separated in polyacrylamide gels and probed with end-labeled antisense oligonucleotides specific for the individual miRNAs indicated. The cellular U6 snRNA served as the loading control.

these viruses for regulating gene expression in cancer cells. Recent studies on small RNAs from infected CEF have identified eight MDV-1-encoded miRNAs that map to the Meq and LAT regions of the genome (8, 9). Furthermore, several recent studies have identified specific miRNA signatures in different tumors, providing insights into the different oncogenic pathways in these tumors (20, 64). In order to analyze the expression profiles of the host- and MDV-1-encoded miRNAs in MD tumor cells, we examined the miRNAs expressed in the MDV-transformed lymphoblastoid cell line MSB-1 by cloning and Northern blot analysis.

One of the conspicuous findings from the analysis of the miRNA sequences from the MSB-1 library was the very large proportion of MDV-1-encoded miRNAs (51%) in relation to the number of host miRNAs (Fig. 4). This level of expression of MDV-1 miRNAs is much higher than that identified from MDV-1-infected CEF, where only 0.6% of the nearly 172,000 reads were miRNAs encoded by MDV-1 (9). The low level of expression of MDV-1 miRNAs in CEF (also evident from the results of the Northern blotting analysis) could partly be explained by the smaller proportions of infected cells in CEF cultures in comparison to MSB-1 cell cultures, where each cell has multiple copies of the MDV genome. However, the increased expression of MDV-1 miRNAs in MSB-1 cells may also be related to the increased lymphocyte-specific expression of these miRNAs in these transformed target cells. An increased proportion of virus-encoded miRNAs over host-encoded miRNAs is not uncommon in transformed cell lines. For example, miRNAs encoded by Kaposi's sarcoma-associated herpesvirus and EBV accounted for >40% of the entire miRNA pool identified from the BC-1 cell line coinfected with these two viruses (11). Once the 518 (10%) MDV-2-encoded miRNAs that we reported previously (66) were also considered, the total proportion of virus-encoded miRNAs in the MSB-1 library was 61%, compared to the 32.2% expression of host-encoded miRNAs. The reasons for the fivefold difference in the levels of miRNAs encoded by the two viruses are not known but may be connected with the differences in relative copy numbers of the two viruses. The precise copy numbers or replication rates of the two viruses in MSB-1 cells are not known. However, on CEF cocultivated with MSB-1 cells, MDV-2 produced 10-fold more plaques than did MDV-1, suggesting that MDV-2 is better adapted for faster replication on CEF (66).

Previous studies of MDV-1-infected CEF identified eight miRNAs that mapped to the Meq and LAT regions of the MDV genome (8, 9). We also cloned all eight of these miRNAs from the MSB-1 library. However, we also identified 5 new MDV-1 miRNAs, taking the total number of MDV-1-encoded miRNAs to 13. As in the case of the eight previously identified miRNAs, the five new MDV-1 miRNAs mapped to the Meq and LAT regions of the genome (Fig. 2). In the MDV genome,

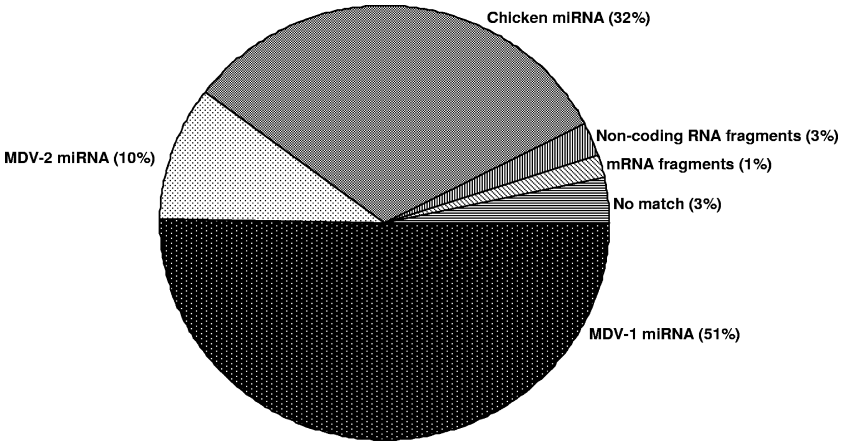


FIG. 4. Pie chart showing proportions of MDV-1, MDV-2, and chicken miRNAs and other molecules cloned from the MSB-1 library.

while most of the genes are transcriptionally silent in latently infected and tumor cells, the repeat (R_L/R_S) regions are generally active (34, 45, 61). Thus, it is not surprising that all of the miRNAs that are expressed at high levels in latently infected/tumor cells are located in a transcriptionally active region of the genome. The genomic locations of the eight previously reported MDV-1-encoded miRNAs have been described (8, 9). Two of the new miRNAs, mdv-miR-M9 and mdv-miR-M12, are also located upstream of the Meq promoter region, like the previously identified miR-M2 to miR-M5 miRNAs, suggesting that these six miRNAs are part of the same transcriptional unit in the same transcriptional orientation as Meq. The high levels of expression of all six of these miRNAs, demonstrated by strong signals in Northern blotting of RNAs from MSB-1 and tumor cells (Fig. 3), suggest that these miRNAs may have major roles in regulating the expression of viral and host genes in latently infected/transformed T cells. The transcription unit of these miRNAs is also antisense to another potential transcript, RLORF8, demonstrated in both CEF and lymphoblastoid cells (52), and hence has the potential to regulate the expression of RLORF8. A recent study demonstrated that EBV-encoded miR-BART2 can downregulate the viral DNA polymerase BALF5 via a similar mechanism (4). However, unlike miR-M2 and miR-M4, which are embedded in the RLORF8 ORF, the newly identified miRNAs miR-M9 and miR-M12 are located downstream of the ORF (Fig. 2).

One of the previously identified miRNAs, miR-M1, mapped downstream of Meq embedded in the ORF of the L1/RLORF5a transcript, although it is not clear whether it affects the expression of this transcript (33, 46, 59). We have identified a new miRNA, mdv-miR-M11, located just downstream of the Meq ORF (Fig. 2). The importance of this novel miRNA is not known, but it is expressed at only very low levels, as indicated by a low cloning frequency in the library and weak signals in Northern blot analysis.

In addition to the three miRNAs, miR-M6 to -M8, that mapped to the LAT region, our studies have revealed miR-M10 and miR-M13, two novel miRNAs encoded from this region. Because these miRNAs are located very close to miR-M6 to -M8 and are in the same transcriptional orientation, these two miRNAs are highly likely to be part of the same cluster. However, compared to miR-M6 to -M8, which are expressed at very high levels (Table 2 and Fig. 3), the levels of expression of miR-M10 and miR-M13, shown by Northern blotting and cloning frequencies, are very low. Although the reasons for their low expression levels are not known, the efficiency in processing of the mature miRNAs could be a factor, especially because of their close proximity within the cluster. For example, miR-M13 is located between the highly expressed miR-M6 and miR-M8 miRNAs, with the mature miRNA sequence of miR-M6 being separated from that of miR-M13 by only a single nucleotide (Table 2). Similarly, the newly identified miR-M10 (only 6 hits in the library) is located just adjacent to the most highly expressed miRNA, miR-M7, which had 800 hits in the library.

Although the expression levels of MDV-1-encoded miRNAs in MSB-1 cells were generally similar to those in tumor tissues, there were clear differences in expression level between infected CEF and transformed MSB-1 cells, with the latter generally expressing higher levels of all miRNAs. However, it was also interesting to see clear differences between miRNAs in the specificity of the strand expressed in infected CEF and MSB-1 cells. The most striking example of strand-specific expression was noted for miR-M7, where the mature miRNA strand, miR-M7-5p, had 800 hits, accounting for 16% of the entire MSB-1 library. Northern blot analysis also revealed very strong expression of this miRNA in both MSB-1 cells and tumor tissues. Although weak signals for miR-M7-5p were detected in the infected CEF, this strand was not identified even once among the nearly 172,000 high-quality reads of small RNA sequences from infected CEF (9), suggesting that this miRNA strand is processed only at very low levels in lytically infected CEF. This cluster of miRNAs maps antisense to the ICP4 gene and to the large intron in the 5′ end of the putative LAT, expressed at high levels in transformed cells/lymphomas as well as in the late stages of lytic infection of CEF (57). Although the reasons for the differences in processing of the two miRNA strands during miRNA biogenesis in this region between CEF and lymphocytes are not clear, it would be interesting to see whether any of the miRNAs play a role in switching between lytic replication and latency. Intriguingly, mdv-miR-M7 also showed evidence of RNA editing. However,

in the absence of knowledge on the targets of any of the MDV-1-encoded miRNAs, the significance of this remains unknown.

Analysis of the miRNA repertoire from MSB-1 cells also identified several host miRNAs, some of which were cloned at high frequencies indicating high levels of expression (Table 1). Some of the more abundant host miRNAs, such as those within the miR-17-92 cluster, have been shown to be amplified in several types of cancer (28, 29). Since these miRNAs have been shown to accelerate the formation of lymphoid malignancies in mouse models (29), the increased expression of these miRNAs in MSB-1 cells could be significant. Similarly, other highly expressed miRNAs, such as miR-21 (249 hits), let-7i (148 hits), the two strands of miR-142 (224 and 243 hits), miR-15a (28 hits), and miR-16 (82 hits), have also been associated with various malignancies, including chronic lymphocytic leukemia (12, 18), suggesting that these miRNAs may contribute toward MDV oncogenicity. Currently, we are examining the roles of different host miRNAs in the induction of lymphomas by MDV. Our studies on MSB-1 cells also revealed six novel chicken miRNAs. The expression of five of these novel miRNAs could be detected by Northern blotting of different chicken tissues, although the expression of miR-454 in all tissues was very weak (Fig. 3b). The expression of both strands of miR-142 appeared to be restricted to the lymphocyte-enriched lungs, thymus, and spleen (Fig. 3b), suggesting that it is likely to be a lymphoid cell-specific miRNA.

We have carried out a study to examine the miRNAome of a herpesvirus-induced lymphoma in chickens by determining the miRNAs expressed in a lymphoblastoid cell line derived from a lymphoma. This study is the first of its kind with an avian lymphoma and has demonstrated that the analysis of the miRNA repertoire would enable investigation of some of the potential pathways used by viruses in neoplastic transformation. A major challenge in the next stage would be the identification of potential targets for some of the miRNAs overexpressed in these cells to identify networks of molecular events regulated by the altered miRNAome in these cells. The present study has not identified miRNAs that are downregulated in transformed cells, whose profiles are also very important for understanding the global events involved in transformation. Currently, we are using microarray analysis of global viral/host miRNA expression in MD tumor cells in relation to that in normal lymphocytes to determine the entire repertoire of upregulated and downregulated miRNAs to identify the extent to which MDV exploits the cellular miRNA pathways to induce neoplastic transformation.

ACKNOWLEDGMENTS

We thank Mihaela Zavolan, Division of Bioinformatics, Biozentrum, University of Basel, Switzerland, for assistance with bioinformatic prediction of MDV-1 miRNAs and Mick Gill for assistance with digital imaging and graphics.

This work was funded by BBSRC, United Kingdom.

REFERENCES

1. **Akiyama, Y., and S. Kato.** 1974. Two cell lines from lymphomas of Marek's disease. Biken J. **17:**105–116.
2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
3. **Bartel, D. P.** 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell **116:**281–297.
4. **Barth, S., T. Pfuhl, A. Mamiani, C. Ehses, K. Roemer, E. Kremmer, C. Jaker, J. Hock, G. Meister, and F. A. Grasser.** 2008. Epstein Barr virus-encoded microRNA miR-BART2 down-regulates the viral DNA polymerase BALF5. Nucleic Acids Res. **36:**666–675.
5. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler.** 2007. GenBank. Nucleic Acids Res. **35:**D21–D25.
6. **Brown, A. C., S. J. Baigent, L. P. Smith, J. P. Chattoo, L. J. Petherbridge, P. Hawes, M. J. Allday, and V. Nair.** 2006. Interaction of MEQ protein and C-terminal-binding protein is critical for induction of lymphomas by Marek's disease virus. Proc. Natl. Acad. Sci. USA **103:**1687–1692.
7. **Burgess, S. C., J. R. Young, B. J. Baaten, L. Hunt, L. N. Ross, M. S. Parcells, P. M. Kumar, C. A. Tregaskes, L. F. Lee, and T. F. Davison.** 2004. Marek's disease is a natural model for lymphomas overexpressing Hodgkin's disease antigen (CD30). Proc. Natl. Acad. Sci. USA **101:**13879–13884.
8. **Burnside, J., E. Bernberg, A. Anderson, C. Lu, B. C. Meyers, P. J. Green, N. Jain, G. Isaacs, and R. W. Morgan.** 2006. Marek's disease virus encodes microRNAs that map to meq and the latency-associated transcript. J. Virol. **80:**8778–8786.
9. **Burnside, J., and R. W. Morgan.** 2007. Genomics and Marek's disease virus. Cytogenet. Genome Res. **117:**376–387.
10. **Buza, J. J., and S. C. Burgess.** 2007. Modeling the proteome of a Marek's disease transformed cell line: a natural animal model for CD30 overexpressing lymphomas. Proteomics **7:**1316–1326.
11. **Cai, X., S. Lu, Z. Zhang, C. M. Gonzalez, B. Damania, and B. R. Cullen.** 2005. Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. Proc. Natl. Acad. Sci. USA **102:**5570–5575.
12. **Calin, G. A., and C. M. Croce.** 2007. Investigation of microRNA alterations in leukemias and lymphomas. Methods Enzymol. **427:**191–213.
13. **Calin, G. A., Y. Pekarsky, and C. M. Croce.** 2007. The role of microRNA and other non-coding RNA in the pathogenesis of chronic lymphocytic leukemia. Best Pract. Res. Clin. Haematol. **20:**425–437.
14. **Calnek, B. W.** 1986. Marek's disease: a model for herpesvirus oncology. CRC Crit. Rev. Microbiol. **12:**293–320.
15. **Cantello, J. L., A. S. Anderson, and R. W. Morgan.** 1994. Identification of latency-associated transcripts that map antisense to the ICP4 homolog gene of Marek's disease virus. J. Virol. **68:**6280–6290.
16. **Chen, I. S., T. W. Mak, J. J. O'Rear, and H. M. Temin.** 1981. Characterization of reticuloendotheliosis virus strain T DNA and isolation of a novel variant of reticuloendotheliosis virus strain T by molecular cloning. J. Virol. **40:**800–811.
17. **Cho, W. C.** 2007. OncomiRs: the discovery and progress of microRNAs in cancers. Mol. Cancer **6:**60.
18. **Cowland, J. B., C. Hother, and K. Gronbaek.** 2007. MicroRNAs and cancer. APMIS **115:**1090–1106.
19. **Cullen, B. R.** 2006. Viruses and microRNAs. Nat. Genet. **38**(Suppl.)**:**S25–S30.
20. **Cummins, J. M., Y. He, R. J. Leary, R. Pagliarini, L. A. Diaz, Jr., T. Sjoblom, O. Barad, Z. Bentwich, A. E. Szafranska, E. Labourier, C. K. Raymond, B. S. Roberts, H. Juhl, K. W. Kinzler, B. Vogelstein, and V. E. Velculescu.** 2006. The colorectal microRNAome. Proc. Natl. Acad. Sci. USA **103:**3687–3692.
21. **Doi, K., A. Kojima, Y. Akiyama, and S. Kato.** 1976. Pathogenicity for chicks of line cells from lymphoma of Marek's disease. Natl. Inst. Anim. Health Q. (Tokyo) **16:**16–24.
22. **Fragnet, L., M. A. Blasco, W. Klapper, and D. Rasschaert.** 2003. The RNA subunit of telomerase is encoded by Marek's disease virus. J. Virol. **77:**5985–5996.
23. **Gimeno, I. M., R. L. Witter, H. D. Hunt, S. M. Reddy, L. F. Lee, and R. F. Silva.** 2005. The pp38 gene of Marek's disease virus (MDV) is necessary for cytolytic infection of B cells and maintenance of the transformed state but not for cytolytic infection of the feather follicle epithelium and horizontal spread of MDV. J. Virol. **79:**4545–4549.
24. **Grey, F., H. Meyers, E. A. White, D. H. Spector, and J. Nelson.** 2007. A human cytomegalovirus-encoded microRNA regulates expression of multiple viral genes involved in replication. PLoS Pathog. **3:**e163.
25. **Griffiths-Jones, S., H. K. Saini, S. van Dongen, and A. J. Enright.** 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res. **36:**D154–D158.
26. **Gupta, A., J. J. Gartner, P. Sethupathy, A. G. Hatzigeorgiou, and N. W. Fraser.** 2006. Anti-apoptotic function of a microRNA encoded by the HSV-1 latency-associated transcript. Nature **442:**82–85.
27. **Hagan, J. P., and C. M. Croce.** 2007. MicroRNAs in carcinogenesis. Cytogenet. Genome Res. **118:**252–259.
28. **Hayashita, Y., H. Osada, Y. Tatematsu, H. Yamada, K. Yanagisawa, S. Tomida, Y. Yatabe, K. Kawahara, Y. Sekido, and T. Takahashi.** 2005. A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. Cancer Res. **65:**9628–9632.
29. **He, L., J. M. Thomson, M. T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S. W. Lowe, G. J. Hannon, and S. M. Hammond.** 2005. A microRNA polycistron as a potential human oncogene. Nature **435:**828–833.
30. **Hirai, K., M. Yamada, Y. Arao, S. Kato, and S. Nii.** 1990. Replicating Marek's disease virus (MDV) serotype 2 DNA with inserted MDV serotype

1 DNA sequences in a Marek's disease lymphoblastoid cell line MSB1-41C. Arch. Virol. **114:**153–165.
31. **ICTVdB Management.** 25 April 2006, posting date. 00.031.1.03. *Mardivirus*. *In* C. Búchen-Osmond (ed.), ICTVdB—The universal virus database, version 4. Columbia University, New York, NY.
32. **Jarosinski, K., L. Kattenhorn, B. Kaufer, H. Ploegh, and N. Osterrieder.** 2007. A herpesvirus ubiquitin-specific protease is critical for efficient T cell lymphoma formation. Proc. Natl. Acad. Sci. USA **104:**20025–20030.
33. **Jarosinski, K. W., N. Osterrieder, V. K. Nair, and K. A. Schat.** 2005. Attenuation of Marek's disease virus by deletion of open reading frame RLORF4 but not RLORF5a. J. Virol. **79:**11647–11659.
34. **Jones, D., L. Lee, J. L. Liu, H. J. Kung, and J. K. Tillotson.** 1992. Marek's disease virus encodes a basic-leucine zipper gene resembling the fos/jun oncogenes that is highly expressed in lymphoblastoid tumors. Proc. Natl. Acad. Sci. USA **89:**4042–4046.
35. **Lee, L. F., K. Nazerian, and J. A. Boezi.** 1975. Marek's disease virus DNA in a chicken lymphoblastoid cell line (MSB-1) and in virus-induced tumours, p. 199–204. *In* G. de Thé, M. A. Epstein, and H. zur Hausen (ed.), Oncogenesis and herpesviruses II. IARC, Lyon, France.
36. **Lee, R. C., and V. Ambros.** 2001. An extensive class of small RNAs in Caenorhabditis elegans. Science **294:**862–864.
37. **Lewis, B. P., C. B. Burge, and D. P. Bartel.** 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell **120:**15–20.
38. **Li, D.-S., J. Pastorek, V. Zelnik, G. D. Smith, and L. J. N. Ross.** 1994. Identification of novel transcripts complementary to the Marek's disease virus homologue of the ICP4 gene of herpes simplex virus. J. Gen. Virol. **75:**1713–1722.
39. **Lupiani, B., L. F. Lee, X. Cui, I. Gimeno, A. Anderson, R. F. Silva, R. L. Witter, H. J. Kung, and S. M. Reddy.** 2004. Marek's disease virus-encoded Meq gene is involved in transformation of lymphocytes but is dispensable for replication. Proc. Natl. Acad. Sci. USA **101:**11815–11820.
40. **Lupiani, B., L. F. Lee, and S. M. Reddy.** 2001. Protein-coding content of the sequence of Marek's disease virus serotype 1. Curr. Top. Microbiol. Immunol. **255:**159–190.
41. **Marek, J.** 1907. Multiple Nervenentzuendung (polyneuritis) bei Huehnern. Dtsch. Tierarztl. Wochenschr. **15:**417–421.
42. **McElroy, J. P., J. C. Dekkers, J. E. Fulton, N. P. O'Sullivan, M. Soller, E. Lipkin, W. Zhang, K. J. Koehler, S. J. Lamont, and H. H. Cheng.** 2005. Microsatellite markers associated with resistance to Marek's disease in commercial layer chickens. Poult. Sci. **84:**1678–1688.
43. **Morgan, R. W., Q. Xie, J. L. Cantello, A. M. Miles, E. L. Bernberg, J. Kent, and A. Anderson.** 2001. Marek's disease virus latency. Curr. Top. Microbiol. Immunol. **255:**223–243.
44. **Nair, V., and H. J. Kung.** 2004. Marek's disease virus oncogenicity: molecular mechanisms, p. 32–48. *In* F. Davison and V. Nair (ed.), Marek's disease, an evolving problem. Elsevier Academic Press, Oxford, United Kingdom.
45. **Ohashi, K., P. H. O'Connell, and K. A. Schat.** 1994. Characterization of Marek's disease virus BamHI-A-specific cDNA clones obtained from a Marek's disease lymphoblastoid cell line. Virology **199:**275–283.
46. **Ohashi, K., W. Zhou, P. H. O'Connell, and K. A. Schat.** 1994. Characterization of a Marek's disease virus BamHI-L-specific cDNA clone obtained from a Marek's disease lymphoblastoid cell line. J. Virol. **68:**1191–1195.
47. **Osterrieder, K., and J. F. Vautherot.** 2004. The genome content of Marek's disease-like viruses, p. 17–31. *In* F. Davison and V. Nair (ed.), Marek's disease, an evolving problem. Elsevier Academic Press, Oxford, United Kingdom.
48. **Osterrieder, N., J. P. Kamil, D. Schumacher, B. K. Tischer, and S. Trapp.** 2006. Marek's disease virus: from miasma to model. Nat. Rev. Microbiol. **4:**283–294.
49. **Parcells, M. S., S. F. Lin, R. L. Dienglewicz, V. Majerciak, D. R. Robinson, H. C. Chen, Z. Wu, G. R. Dubyak, P. Brunovskis, H. D. Hunt, L. F. Lee, and H. J. Kung.** 2001. Marek's disease virus (MDV) encodes an interleukin-8 homolog (vIL-8): characterization of the vIL-8 protein and a vIL-8 deletion mutant MDV. J. Virol. **75:**5159–5173.
50. **Payne, L. N., K. Howes, M. Rennie, J. M. Bumstead, and A. W. Kidd.** 1981. Use of an agar culture technique for establishing lymphoid cell lines from Marek's disease lymphomas. Int. J. Cancer **28:**757–766.
51. **Peng, Q., and Y. Shirazi.** 1996. Characterization of the protein product encoded by a splicing variant of the Marek's disease virus Eco-Q gene (Meq). Virology **226:**77–82.
52. **Peng, Q., and Y. Shirazi.** 1996. Isolation and characterization of Marek's disease virus (MDV) cDNAs from a MDV-transformed lymphoblastoid cell line: identification of an open reading frame antisense to the MDV Eco-Q protein (Meq). Virology **221:**368–374.
53. **Pfeffer, S., M. Zavolan, F. A. Grasser, M. Chien, J. J. Russo, J. Ju, B. John, A. J. Enright, D. Marks, C. Sander, and T. Tuschl.** 2004. Identification of virus-encoded microRNAs. Science **304:**734–736.
54. **Ramkissoon, S. H., L. A. Mainwaring, Y. Ogasawara, K. Keyvanfar, J. P. McCoy, Jr., E. M. Sloand, S. Kajigaya, and N. S. Young.** 2006. Hematopoietic-specific microRNA expression in human cells. Leukoc. Res. **30:**643–647.
55. **Reese, S., G. Dalamani, and B. Kaspers.** 2006. The avian lung-associated immune system: a review. Vet. Res. **37:**311–324.
56. **Rhiza, H.-J., and B. Bauer.** 1984. Persistence of viral DNA in Marek's disease virus-transformed lymphoblastoid cell lines, p. 481–488. *In* G. Wittman, R. M. Gaskell, and H.-J. Rhiza (ed.), Latent herpesvirus infections in veterinary medicine. Martinus Nijhoff, Boston, MA.
57. **Ross, N. L.** 1999. T-cell transformation by Marek's disease virus. Trends Microbiol. **7:**22–29.
58. **Schat, K. A., B. W. Calnek, and J. Fabricant.** 1982. Characterisation of two highly oncogenic strains of Marek's disease virus. Avian Pathol. **11:**593–605.
59. **Schat, K. A., B. J. Hooft van Iddekinge, H. Boerrigter, P. H. O'Connell, and G. Koch.** 1998. Open reading frame L1 of Marek's disease herpesvirus is not essential for in vitro and in vivo virus replication and establishment of latency. J. Gen. Virol. **79:**841–849.
60. **Subramanian, S., W. O. Lui, C. H. Lee, I. Espinosa, T. O. Nielsen, M. C. Heinrich, C. L. Corless, A. Z. Fire, and M. van de Rijn.** 8 October 2007. MicroRNA expression signature of human sarcomas. Oncogene. doi: 10.1038/sj.onc.1210836.
61. **Sugaya, K., G. Bradley, M. Nonoyama, and A. Tanaka.** 1990. Latent transcripts of Marek's disease virus are clustered in the short and long repeat regions. J. Virol. **64:**5773–5782.
62. **Takagi, M., T. Takeda, Y. Asada, C. Sugimoto, M. Onuma, and K. Ohashi.** 2006. The presence of a short form of p53 in chicken lymphoblastoid cell lines during apoptosis. J. Vet. Med. Sci. **68:**561–566.
63. **Trapp, S., M. S. Parcells, J. P. Kamil, D. Schumacher, B. K. Tischer, P. M. Kumar, V. K. Nair, and N. Osterrieder.** 2006. A virus-encoded telomerase RNA promotes malignant T cell lymphomagenesis. J. Exp. Med. **203:**1307–1317.
64. **Volinia, S., G. A. Calin, C. G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. Harris, and C. M. Croce.** 2006. A microRNA expression signature of human solid tumors defines cancer gene targets. Proc. Natl. Acad. Sci. USA **103:**2257–2261.
65. **Xu, H., X. Wang, Z. Du, and N. Li.** 2006. Identification of microRNAs from different tissues of chicken embryo and adult chicken. FEBS Lett. **580:**3610–3616.
66. **Yao, Y., Y. Zhao, H. Xu, L. P. Smith, C. H. Lawrie, A. Sewer, M. Zavolan, and V. Nair.** 2007. Marek's disease virus type 2 (MDV-2)-encoded microRNAs show no sequence conservation with those encoded by MDV-1. J. Virol. **81:**7164–7170.
67. **Zanette, D. L., F. Rivadavia, G. A. Molfetta, F. G. Barbuzano, R. Proto-Siqueira, and W. A. Silva, Jr.** 2007. miRNA expression profiles in chronic lymphocytic and acute lymphocytic leukemia. Braz. J. Med. Biol. Res. **40:**1435–1440.
68. **Zuker, M.** 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31:**3406–3415.
69. **zur Hausen, H.** 2001. Oncogenic DNA viruses. Oncogene **20:**7820–7823.

Journal of General Virology (2009), **90**, 1551–1559 DOI 10.1099/vir.0.009902-0

Differential expression of microRNAs in Marek's disease virus-transformed T-lymphoma cell lines

Yongxiu Yao,[1] Yuguang Zhao,[1] Lorraine P. Smith,[1] Charles H. Lawrie,[2] Nigel J. Saunders,[3] Michael Watson[1] and Venugopal Nair[1]

Correspondence
Venugopal Nair
venu.gopal@bbsrc.ac.uk

[1]Division of Microbiology, Institute for Animal Health, Compton RG20 7NN, UK

[2]LRF Molecular Haematology Unit, Nuffield Department of Clinical Laboratory Sciences, University of Oxford, Oxford OX3 9DU, UK

[3]Bacterial Pathogenesis and Functional Genomics Group, Sir William Dunn School of Pathology, University of Oxford, Oxford OX1 3RE, UK

MicroRNAs (miRNAs) are increasingly recognized to play crucial roles in regulation of gene expression in different biological events, including many sporadic forms of cancer. However, despite the involvement of several viruses in inducing cancer, only a limited number of studies have been carried out to examine the miRNA expression signatures in virus-induced neoplasia, particularly in herpesvirus-induced tumours where virus-encoded miRNAs also contribute significantly to the miRNome of the tumour cell. Marek's disease (MD) is a naturally occurring, rapid-onset $CD4^+$ T-cell lymphoma of poultry, induced by the highly contagious Marek's disease virus (MDV). High levels of expression of virus-encoded miRNAs and altered expression of several host-encoded miRNAs were demonstrated in the MDV-transformed lymphoblastoid cell line MSB-1. In order to identify the miRNA expression signature specific to MDV-transformed cells, we examined the global miRNA expression profiles in seven distinct MDV-transformed cell lines by microarray analysis. This study revealed that, in addition to the high levels of MDV-encoded miRNAs, these MD tumour-derived lymphoblastoid cell lines showed altered expression of several host-encoded miRNAs. Comparison of the miRNA expression profiles of these cell lines with the MDV-negative, retrovirus-transformed AVOL-1 cell line showed that miR-150 and miR-223 are downregulated irrespective of the viral aetiology, whereas downregulation of miR-155 was specific for MDV-transformed tumour cells. Thus, increased expression of MDV-encoded miRNAs with specific downregulation of miR-155 can be considered as unique expression signatures for MD tumour cells. Analysis of the functional targets of these miRNAs would contribute to the understanding of the molecular pathways of MD oncogenicity.

Received 22 December 2008
Accepted 15 March 2009

INTRODUCTION

MicroRNAs (miRNAs) are small (approx. 22–25 nt long), non-coding RNAs that regulate gene expression by base pairing with the RNA transcripts, targeting them for translational repression or degradation. All metazoan genomes encode miRNAs and the latest release (12.0) of miRBase (http://microrna.sanger.ac.uk) contains 8619 hairpin precursor miRNAs in various species (Griffiths-Jones *et al.*, 2008). In the past few years, there have been huge increases in the number of studies on miRNA and cancer. Profiling of global miRNA expression levels (miRNome) has generated extensive data on miRNA expression in various forms of cancer. These studies have reiterated the important role of miRNAs in all aspects of cancer biology, including proliferation, apoptosis, invasion/metastasis and angiogenesis (Fabbri *et al.*, 2008; Lee & Dutta, 2009). Such studies have also provided information on the developmental lineage, differentiation state and prognosis of malignant cells (Lowery *et al.*, 2008; Schotte *et al.*, 2008). Nearly all of these studies have been carried out on non-infectious forms of cancer. Current estimates suggest that viruses are involved in 15–20 % of human cancers worldwide (Javier & Butel, 2008) and oncogenic viruses have been instrumental in delineating several molecular pathways of neoplastic transformation. Despite this, comparatively little is known on global miRNA expression profiles of virus-induced cancers (Martinez *et al.*, 2008; Yeung *et al.*, 2008). In many tumours, particularly those associated with oncogenic herpesviruses (Cosmopoulos *et al.*, 2008; Cullen, 2006; Gottwein & Cullen, 2008; Pfeffer *et al.*, 2005; Sullivan & Grundhoff, 2007), high levels of expression of virus-encoded miRNAs

Supplementary figures and tables are available with the online version of this paper.

add further complexity to the miRNome of the transformed cell (Ghosh *et al.*, 2008).

Marek's disease virus (MDV) is a highly contagious, oncogenic alphaherpesvirus of the genus *Mardivirus*; it is associated with Marek's disease (MD), a naturally occurring, rapid-onset T-cell lymphoma of chicken (Calnek, 1986). The MDV genome encodes several miRNAs that map to the MDV-encoded oncogene *Meq* and the LAT (latency-associated transcript) regions of the virus (Burnside *et al.*, 2006; Burnside & Morgan, 2007). Although the target genes regulated by MDV-encoded miRNAs are yet to be discovered, high levels of their expression in MDV-transformed cell lines and tumours suggest that they play important roles in oncogenesis (Morgan *et al.*, 2008; Xu *et al.*, 2008). From a small RNA library generated from MDV-transformed lymphoblastoid cell line MSB-1 (Akiyama & Kato, 1974), we have previously demonstrated high levels of expression of MDV-encoded miRNAs (Yao *et al.*, 2007, 2008). Elevated levels of expression of some of these miRNAs were also confirmed by real-time quantitative PCR in these cells (Xu *et al.*, 2008). Several host miRNAs that are associated directly with oncogenicity, such as miR-17-92, miR-21 and let-7i, were also present at a high frequency in the MSB-1 library, suggesting a role for these miRNAs in neoplastic transformation of these cells (Yao *et al.*, 2008). Although analysis of miRNAs by cloning (Yao *et al.*, 2008) or high-throughput sequencing (Burnside *et al.*, 2008) is used to identify upregulated miRNAs in tumours or transformed cells, such studies do not provide differential expression profiles of miRNAs in different cell types. Comparisons of miRNA expression profiles between neoplastically transformed and normal cells using miRNA microarrays have enabled the identification of specific miRNA expression signatures in different types of cancer cell, some of which have been shown to be useful indicators of cell type, stage of differentiation and even prognosis of the cancer (Calin & Croce, 2006; Rosenfeld *et al.*, 2008). Only a limited number of studies comparing the global miRNA expression profiles of virally transformed tumour cells have been carried out, particularly in tumour cells transformed by oncogenic herpesviruses that themselves encode multiple miRNAs (Sullivan & Grundhoff, 2007). Here we describe the results of the comparison of the miRNA expression profile of seven different MDV-transformed T-lymphoblastoid cell lines with that of normal chicken splenocyte or $CD4^+$ T-cell populations.

METHODS

Transformed cell lines. Small RNA prepared from seven independent $CD4^+$ T-lymphoma cell lines derived from MDV-1-induced tumours was used for miRNA expression profiling. The cell lines studied are MDCC-MSB1 from a spleen lymphoma induced by the BC-1 strain of MDV-1 (Akiyama & Kato, 1974), MDCC-HP8 from a GA strain-induced tumour (Nazerian, 1987) and five cell lines (MDCC-226S, MDCC-265L, MDCC-273S, MDCC-299K and MDCC-299L) established from lymphomas of birds infected with RB-1B virus (Petherbridge *et al.*, 2004). Reticuloendotheliosis virus T (REV-T strain)-transformed $CD4^+$ T-cell line AVOL-1 (Yao *et al.*, 2008) and avian leukosis virus (ALV) HPRS F42 strain-transformed B-cell line HP45 (Nazerian, 1987) were included in the experiments as MDV-negative transformed cell lines. Cell lines were grown at 38.5 °C in 5 % CO_2 in RPMI 1640 medium containing 10 % fetal calf serum, 10 % tryptose phosphate broth and 1 % sodium pyruvate.

Chicken splenocytes, $CD4^+$ T cells and magnetic cell sorting. Single-cell suspensions of lymphocytes were prepared from spleen tissues of uninfected birds by using Histopaque-1083 (Sigma-Aldrich) density-gradient centrifugation. $CD4^+$ T cells were isolated by magnetic cell sorting using mouse anti-chicken CD4 antibodies (Chan *et al.*, 1988) and goat anti-mouse IgG microbeads (Miltenyi Biotec). After each antibody treatment, cells were washed three times with PBS containing 0.5 % bovine serum albumin. At each wash, the cell suspension was centrifuged at 450 ***g*** for 10 min. Positively stained cells were sorted through an AutoMACS Pro Separator (Miltenyi Biotec). Purity of the sorted cells was confirmed to be >99 % by flow cytometry after labelling with monoclonal anti-goat/sheep IgG–fluorescein isothiocyanate (Sigma) antibody (data not shown).

Microarray analysis of miRNA expression. Preparation of probes and hybridization to the arrays were carried out by using methods described previously (Lawrie *et al.*, 2007, 2008). Briefly, 500 ng purified miRNA from lymphoblastoid cell lines, normal splenocytes or $CD4^+$ T-cell populations was labelled with either Cy3 or Cy5 dye using an Array 900microRNA RT kit from Genisphere and hybridized to the μRNA microarray described previously (Lawrie *et al.*, 2008). The array contains miRNA probe sets (designed from miRBase v. 9.2) spotted in quadruplicate non-adjacently (the sequences of the probes are available at http://www.microRNAworld.com). Where the homologous sequences of miRNAs in different species are identical, miRNA sequences from only one species was spotted on the array. In addition, probes for the mature MDV-1-encoded miRNAs miR-M2-3p, miR-M2-5p, miR-M3-3p, miR-M3-5p, miR-M4-3p, miR-M4-5p, miR-M5-3p, miR-M11-5p and miR-M12-3p (Yao *et al.*, 2008) were also included on the array. A model design with splenocytes and/or $CD4^+$ T cells as reference was used and the expression values are depicted as log ratios of test and reference samples. Image analysis was done by using BlueFuse software (BlueGnome).

Statistical analysis of microarray data. Data were normalized within each microarray by subtracting the mean $\log_2$ ratio from each measurement. Quantile normalization was then used to standardize the data across arrays, and a linear model was fitted to each miRNA using Limma (Smyth, 2005). The resultant *P*-values were adjusted for multiple testing by using the Benjamini–Hochberg correction of the false-discovery rate (Benjamini & Hochberg, 1995). The *P*-values and mean expression were calculated for each cell type. Those miRNAs showing consistent differential expression across all cell types were subjected to a second analysis where all samples were treated as 'virus-infected', regardless of cell type. A similar analysis in Limma was performed, resulting in *P*-values calculated for expression across all cell types. The data were sorted according to the $\log_2$ ratio and heat maps were produced by using R (http://www.R-project.org).

Inducible expression of mature miRNAs. Constructs for inducible expression of miRNAs were generated in the pRTS-1-SVP-Tom(−) vector, constructed from the pRTS-1 plasmid (Bornkamm *et al.*, 2005) by replacing the hygromycin B-resistance gene with a puromycin-resistance gene. The inducible bidirectional promoter in this construct expresses monomeric red fluorescent protein *td*-tomato (Shaner *et al.*, 2004) and the miRNA of interest simultaneously in a tightly regulated, doxycycline (Dox)-inducible system. The inserted sequence of each miRNA consists of the stem–loop structure and 100–200 bp of upstream and downstream flanking genome sequences, a feature

designed to ensure that the expressed miRNAs are processed as naturally as possible. The approximately 500 bp fragment of each miRNA was obtained by RT-PCR using RNA extracted from chicken embryo fibroblasts by using the following primers: gga-miR-155 (5′-AGATCTCTGATGTCTGTACTCTTTATGAC-3′ and 5′-CTCGAGCC-CAGTGCCCTTAACTTAG-3′); gga-miR-223 (5′-AGATCTGCAA-CGTCTGTCCTGTCC-3′ and 5′-CTCGAGCAGGAACTGTACC-AGCAG-3′). As the sequence of the chicken orthologue was not available, we used hsa-miR-150 for generating the expression constructs of miR-150, using RNA prepared from HEK293T cells with primers 5′-AGATCTCTTCCTGCCCTCTTTGATG-3′ and 5′-CTCGAGCA-ATAGAAACAGGTGTACTTTG-3′. More recently, the mature gga-mir-150 sequence was shown to be identical to the hsa-miR-150 sequence except for a single nucleotide substitution (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6541). Agarose gel-purified PCR fragments were cloned into the *Bgl*II and *Xho*I sites of the pRTS-1-SVP-Tom(−) vector and confirmed by sequence analysis. For the expression of cloned miRNAs, HEK293T cells were transfected with the pRTS-1–miRNA constructs by using Lipofectamine 2000 (Invitrogen) and stable cells were selected with puromycin (1 μg ml^{-1}). Respective miRNAs induced with the addition of Dox (1 μg ml^{-1}) were detected by Northern blotting analysis.

Luciferase reporter assays. In order to demonstrate that the Dox-induced miRNAs in stably transfected HEK293T cells are functional, we carried out reporter assays on one of the stable cell lines expressing gga-miR-155 by analysing its effect on the previously characterized target Pu.1. For this, we generated a reporter construct in which a 110 bp fragment of the chicken Pu.1 3′ untranslated region (UTR) transcript (GenBank accession no. NM_205023) that contained the predicted miR-155-response element (MRE) is inserted downstream of *Renilla* luciferase in the psiCHECK-2 vector (Promega) to generate the Pu.1-3′UTR-wt construct. A mutant construct (Pu.1-3′UTR-mt) with mutations in the MRE sequences was also constructed. The details of the oligonucleotides used for generating the chicken Pu.1-3′UTR reporter constructs are shown elsewhere (Zhao *et al.*, 2009). HEK293T cells stably expressing gga-miR-223 were used as a control. Transfection into HEK293T cells for luciferase reporter assays was carried out in 96-well plates with Lipofectamine 2000 (Invitrogen). Firefly and *Renilla* luciferase activities were measured consecutively with the Dual-Luciferase Reporter Assay system (Promega), using a Lucy-1 luminometer (Anthos Labtec). In all cases, the constitutively expressed firefly luciferase activity in the psiCHECK-2 vector served as a normalization control for transfection efficiency.

Preparation of RNA for Northern blotting. Preparation of RNA for Northern blotting analysis from splenocytes, CD4$^+$ and CD4$^-$ populations of T lymphocytes, MDV- and retrovirus-transformed cell lines, and HEK293T cells stably selected for recombinant pRTS-1 vectors expressing miRNAs was carried out by using standard methods as described previously (Yao *et al.*, 2007, 2008). Antisense oligonucleotides to the miRNAs gga-miR-155, gga-miR-223 and hsa-miR-150 were used as probes.

RESULTS

In order to determine the miRNA expression signature in MDV-induced tumours, we compared the global gene expression profiles of seven MDV-transformed cell lines by microarray analysis using miRNA probe sets designed from miRBase v. 9.2. The tests validating the sensitivity, specificity and reproducibility of these arrays have been described previously (Lawrie *et al.*, 2008). Ratios of miRNA expression levels in transformed cell lines normalized to the reference samples of normal chicken splenocytes or CD4$^+$ cells were used for the analysis. These studies showed that the grouping of miRNAs in the seven MDV-1-transformed cell lines was distinct from that in the MDV-1-negative lymphocyte cell line AVOL-1 when splenocytes were used as a reference (Fig. 1a). When purified CD4$^+$ T cells were used as a reference, the expression profiles of miRNAs in four of these cell lines (Fig. 1b) were largely in agreement with those obtained by using reference splenocytes (detailed data on the $\log_2$ fold changes and the *P*-values for each of the cell lines are provided in Supplementary Tables S1 and S2, available in JGV Online). The miRNA profiles of the cell lines in both of these analyses were generally consistent, although individual cell lines did show differences (Supplementary Figs S1 and S2).

MDV-1-encoded miRNAs are upregulated in transformed cell lines

The inclusion of mature probe sequences of nine MDV-1-encoded miRNA sequences (Yao *et al.*, 2008) alongside the host miRNA probe sets in the miRNA microarray enabled assessment of the levels of expression of virus- and host-encoded miRNAs in these cell lines. Compared with the MDV-negative REV-T-transformed cell line AVOL-1, all of the MDV-transformed cell lines showed upregulation of MDV-1-encoded miRNAs (Fig. 1a), although the expression levels of individual miRNAs were not uniform in these cells.

Changes in host miRNA profiles in MD tumour cell lines

Examination of the global miRNA expression profiles of the seven MDV-1-transformed cell lines revealed changes in several host miRNAs (Fig. 1). Major differences in the host miRNA expression profiles could also be observed between MDV-1-transformed cell lines and the MDV-negative cell line AVOL-1, demonstrating the differences in the molecular oncogenic pathways between these cell lines. Microarray readouts from our studies did demonstrate differences between cell lines with regard to the expression of individual miRNAs. Although such differences could be important for individual cell lines, our main interest was to look for miRNA profiles that are conserved in all MDV cell lines, as this could give insights into the fundamental molecular pathways of miRNA-mediated gene regulation in MDV transformation. Our results showed that several host-encoded miRNAs, such as miR-155, miR-223, miR-150, miR-451, miR-26a and miR-126, were downregulated in all MDV-transformed cell lines relative to the levels in normal splenocytes or CD4$^+$ T cells (Fig. 1). As miR-223 and miR-150 were also downregulated in retrovirus-transformed AVOL-1 cells, the reduced expression of these two miRNAs is thought to be a broader feature of transformed T cells, irrespective of the viral aetiology. However, this was not the case with miR-155, the levels of which were consistently reduced in all of the seven

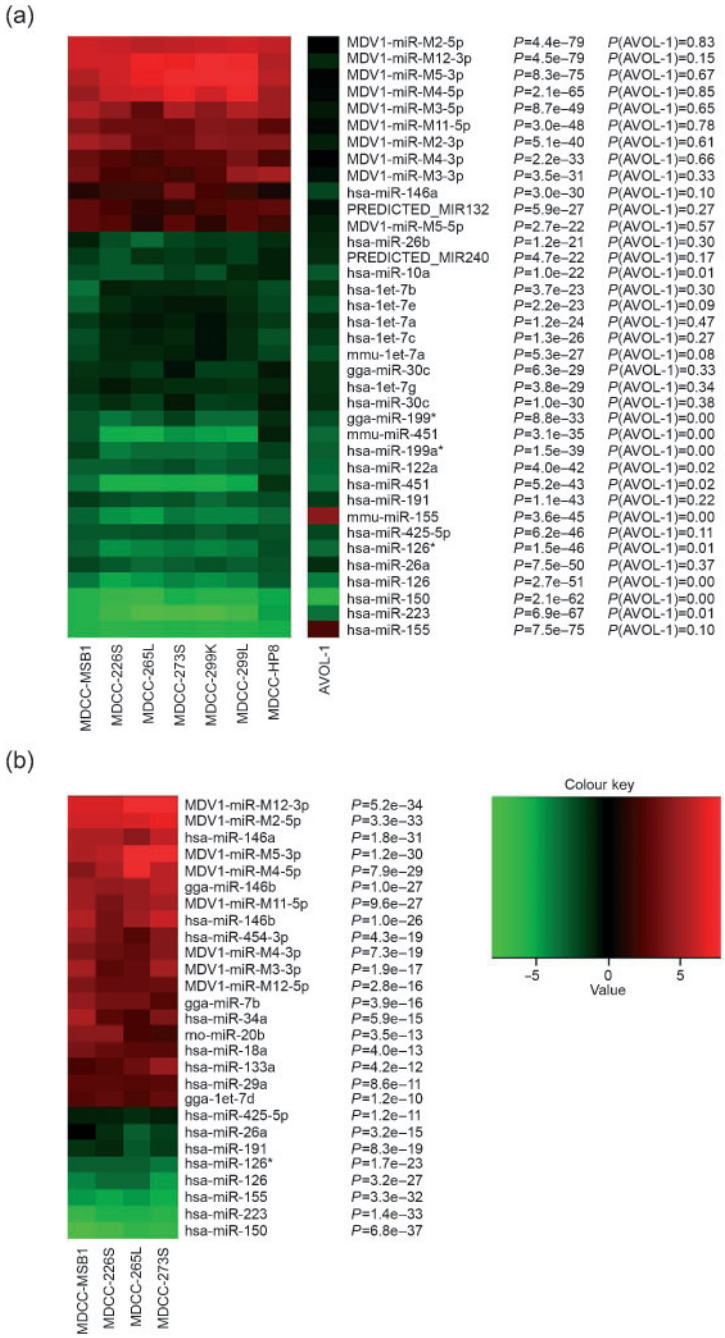


Fig. 1. Heat maps showing clustering of differentially expressed miRNAs (adjusted $P<0.05$) in MDV-transformed cell lines. (a) miRNAs identified in the seven MDV-transformed cell lines (names of cell lines are shown below each lane) in comparison to those in the MDV-1-negative REV-transformed AVOL-1 cell line. The data shown are normalized by using uninfected splenocytes as the reference. (b) Heat map showing differentially expressed miRNAs in four of the above MDV-transformed cell lines (shown below each lane), normalized against expression in normal $CD4^+$ T cells as the reference. A colour key indicating low (green) to high (red) values and the *P*-values is also shown.

MDV-1-transformed lymphoblastoid cell lines whilst the expression levels in the AVOL-1 cell line were very high (Fig. 1a), demonstrating that the downregulation of miR-155 is a feature unique to MDV transformation of T cells.

As all of the MDV-1-transformed cell lines used in this study have a $CD4^+$ T-cell phenotype, we also wondered whether it is appropriate to use whole splenocyte populations as the reference sample in the analysis. In order to rule out the possibility that the altered expression profiles of miRNAs in the MDV-transformed cell lines are not due to the use of splenocytes as the reference, we also repeated the analysis of miRNA expression in four MDV-1-transformed cell lines with purified $CD4^+$ T cells as the reference. Overall, the results were largely consistent with the values obtained by using normal splenocytes as the reference (Fig. 1b). However, the use of purified $CD4^+$ cells as the reference resulted in the demonstration of increased expression of several more host miRNAs, such as miR-146b, miR-454, miR-7b, miR-34a, miR-18a, miR-133a, miR-29a and gga-let-7d (Fig. 1b). As the splenocyte reference data represent the miRNA expression of multiple

lymphocyte populations, it was not surprising to see changes in the miRNA expression patterns when purified $CD4^+$ T cells were used as the reference. Data on the $\log_2$ fold changes and *P*-values for each cell line using splenocytes or purified $CD4^+$ T cells as the reference can be seen in Supplementary Tables S1 and S2, respectively.

Analysis of miRNA expression by Northern blotting

For further validation of the microarray data demonstrating the reduced expression of miR-155, miR-223 and miR-150 in MDV-transformed tumour cell lines, we carried out Northern blotting analysis comparing the levels of these miRNAs in four MDV-transformed cell lines with those in normal lymphocyte populations and retrovirus-transformed cell lines AVOL-1 and HP45. Northern blotting analysis confirmed the observations of reduced expression of these three miRNAs in the microarray readouts of MDV-transformed cells. The levels of gga-miR-155 signal detected were very low in the normal splenocyte and $CD4^+$/$CD4^-$ T-cell populations (Fig. 2a). These low levels were reduced further in all MDV-transformed cell lines included in the assay. In contrast, stronger gga-miR-155 signals were evident in avian retrovirus-transformed cell lines HP45 and AVOL-1. The levels of gga-miR-223 were lower in normal $CD4^+$ T cells than in whole spleen-cell or $CD4^-$ T-cell populations. However, no hybridization signals for gga-miR-223 expression were evident in any of the cell lines transformed by either MDV-1 or avian retroviruses, suggesting that the downregulation of gga-miR-223 is a broader feature of lymphocyte transformation, irrespective of the viral aetiology. The expression of miR-150 also appeared to be restricted to the untransformed cells, as there was no evidence of its expression in transformed cells. Although miR-150 and miR-223 expression appeared to be similar in this respect, the levels of miR-150 were much higher in $CD4^+$ T cells (Fig. 2a). We also evaluated the Dox-inducible expression of miR-155, miR-223 and miR-150 from the pRTS-1 vector (Bornkamm *et al.*, 2005). Northern blotting analysis of HEK293T cells expressing the pRTS-1–miRNA constructs showed high levels of expression of each of the three mature miRNAs, regulated tightly in a Dox-inducible manner (Fig. 2b).

For functional evaluation of the efficacy of this inducible expression system for identifying potential miRNA targets, we examined the ability of the pRTS-1–miR-155 expression vector to silence the reporter construct containing the wild-type or the mutant MRE region of the 3′ UTR of Pu.1, a validated target of miR-155 (Zhao *et al.*, 2009). This assay showed that the relative *Renilla* luciferase levels of reporter constructs with wild-type MRE sequences were reduced specifically by nearly 60 % compared with the mutant MRE construct (Fig. 3a). This reduction in luciferase levels was dependent on the induction of miR-155 in these cells by Dox treatment. The specificity of the reporter assay was

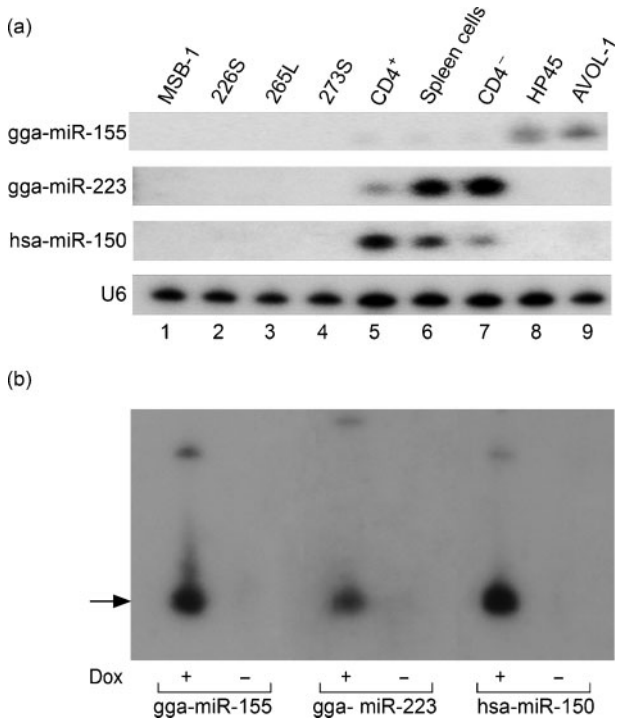


Fig. 2. Northern blotting analysis of differentially expressed miRNAs. (a) Twenty micrograms of total RNA extracted from the indicated cells was separated on a 15 % denaturing polyacrylamide gel, blotted and hybridized with end-labelled antisense oligonucleotide probes to gga-miR-155, gga-miR-223 and hsa-miR-150. The cellular U6 small nuclear RNA served as the loading control. (b) Dox-inducible expression of miRNAs from pRTS-1 constructs in HEK293T cells. Mature miRNAs (indicated by an arrow) in Dox+ samples are shown. Absence of signals in the untreated (Dox−) lanes demonstrates the specificity of the miRNA inducible system. Bands representing each of the three pre-miRNAs can also be seen.

demonstrated further by the absence of reduction in luciferase levels in cells expressing gga-miR-223 (Fig. 3a). The tightly regulated nature of the Dox-inducible expression system used here was demonstrated by the non-leaky expression of the *td*-tomato marker gene in the untreated (Dox−) cells (Fig. 3b).

DISCUSSION

Global changes in miRNA expression profiles using microarray analysis are used increasingly to identify specific miRNA expression signatures associated with several human malignancies (Calin & Croce, 2006, 2007; Lawrie *et al.*, 2008). Most of these studies have been carried out on tumour tissues or cell lines derived from various sporadic forms of cancer (Ozen *et al.*, 2008; Ruike *et al.*, 2008). These studies have highlighted the direct oncogenic potential of the cluster of miRNAs such as miR-21, miR-155 and miR-17-92, providing valuable insights into the

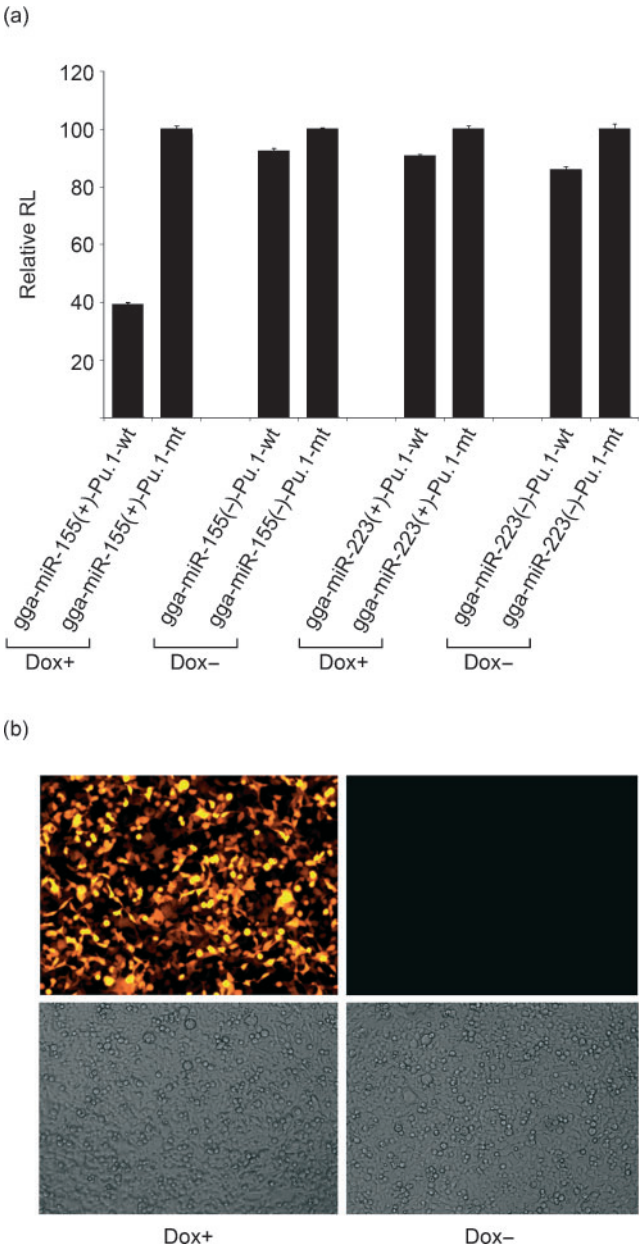


Fig. 3. Functional analysis of inducible miR-155 by using a reporter assay on the Pu.1 3′ UTR. (a) Reporter assays in HEK293T cells expressing gga-miR-155 and gga-miR-223 from the Dox-inducible promoter in the pRTS-1 vector. Relative *Renilla* luciferase levels (RL) of Dox+ and Dox− HEK293T cells transfected with luciferase reporter constructs containing the wild-type (Pu.1-wt) or mutant (Pu.1-mt) MRE-containing region of the 3′ UTR of the chicken Pu.1 transcript (Zhao *et al.*, 2009) are shown. (b) Fluorescence microscopic image of *td*-tomato expression in Dox+ (top left) and Dox− (top right) HEK293T cells stably expressing pRTS-1−gga-miR-155. Bright-field images of the two cell populations are also shown (bottom left and right).

molecular pathways of oncogenesis (Wiemer, 2007). Oncogenic viruses account for a large proportion of neoplasms in man and animals (Javier & Butel, 2008). Although the induction of many of these tumours has until recently been attributed mainly to virus-encoded oncoproteins, an increasing amount of data indicates that miRNAs, encoded either by the host or by viruses themselves in the case of oncogenic herpesviruses (Cullen, 2006, 2009; Pfeffer *et al.*, 2004, 2005; Sullivan & Grundhoff, 2007), play significant roles in oncogenesis.

We and others have documented recently that the highly oncogenic MDV-1 encodes several novel miRNAs (Burnside & Morgan, 2007; Burnside *et al.*, 2008; Morgan *et al.*, 2008; Yao *et al.*, 2008). High levels of expression of these miRNAs in lymphomas and transformed cell lines have been demonstrated by direct cloning, Northern blotting and quantitative RT-PCR (Burnside *et al.*, 2006; Xu *et al.*, 2008; Yao *et al.*, 2008). Although these studies have been valuable in identifying miRNAs that are expressed at high levels in these cells, they do not always provide comprehensive miRNA expression profiles, particularly of those miRNAs that are downregulated in the transformed cells. With a view to examining the global expression of miRNAs in MDV-transformed cells, we carried out miRNA expression profiling of seven independent MDV-transformed tumour cell lines by using miRNA microarray analysis. Each of these cell lines was derived from an independent MD lymphoma. As demonstrated previously with MSB-1 cells (Yao *et al.*, 2007, 2008), MDV-1-encoded miRNAs were indeed the most abundant miRNAs in all of the cell lines. High-level expression of virus-encoded miRNAs appears to be a feature common to virus-transformed cell lines, as cells transformed by other oncogenic herpesviruses, such as Kaposi's sarcoma-associated herpesvirus (KSHV) and Epstein–Barr virus, also showed high levels of expression of virus-encoded miRNAs (Cai *et al.*, 2005; Lawrie *et al.*, 2008; Pratt *et al.*, 2009). Higher copy numbers of viral genomes and active transcription of miRNA genes may account for the higher expression of virus-encoded miRNAs in the virus-transformed cell lines, although one cannot rule out the possibility of differential processing of virus-encoded miRNAs in these cells. The functions and putative targets of most of the MDV-encoded miRNAs remain unknown. However, we have shown recently that MDV-miR-M4, one of the most abundantly expressed virus-encoded miRNAs in all of the cell lines, is a functional orthologue of miR-155 with the potential to target important lymphocyte-specific transcription factors such as Pu.1 (Zhao *et al.*, 2009). More efforts in the future to identify the potential targets of other MDV-encoded miRNAs (Morgan *et al.*, 2008) will unravel more molecular pathways of oncogenesis.

Microarray data analysis of the changes in the global expression profiles of host-encoded miRNAs in MDV-transformed cell lines could be grouped into (i) those that are restricted only to some of the MDV-transformed cell lines and (ii) those that appear to be conserved across all of the cell lines. The changes in miRNA expression in individual cell lines, such as the increased expression of miR-221/miR-222 in MSB-1 cells (Lambeth *et al.*, 2009),

are likely to be important in the regulation of respective target proteins in individual cell lines. However, in this paper, we focus on the global changes in miRNA expression common to all MDV-transformed cell lines.

The miRNA profile of the seven MD tumour cell lines showed changes in the expression of several miRNAs. These included the downregulation of miR-150, miR-223 and miR-155, confirmed by Northern blotting analyses (Figs 1 and 2a). Of these, miR-150 and miR-223 were also downregulated in AVOL-1 cells, demonstrating it to be a broader feature of lymphocyte transformation, regardless of the viral aetiology. Reduced expression of miR-150 has also been reported in human lymphoid malignancies such as diffuse large B cell lymphoma (Garzon & Croce, 2008; Landgraf *et al.*, 2007; Lawrie *et al.*, 2008), indicating its conserved function across different species. Increasing evidence suggest that miR-150 functions through the transcription factor c-*myb* (Garcia & Frampton, 2008; Lin *et al.*, 2008; Lu *et al.*, 2008; Xiao *et al.*, 2007; Zhou *et al.*, 2007), and the dysregulation of c-*myb* and its targets could be important in T-cell transformation. In the case of miR-223, although all of the regulatory mechanisms are not yet understood fully, recent studies have indicated a clear role for miR-223 in haematopoiesis, as well as in malignancies (Baek *et al.*, 2008; Garzon & Croce, 2008; Johnnidis *et al.*, 2008; Merkerova *et al.*, 2008). Whilst identification of the potential targets is important to understand fully the molecular pathways, involvement of miR-223 appears to be logical in MDV-induced lymphocyte transformation.

The downregulation of miR-155 observed by microarray analysis was unique to MDV-transformed cell lines, as it was upregulated in the MDV-negative AVOL-1 cell line (Fig. 1). Although the levels of miR-155 in the normal lymphocyte populations were low by Northern blotting analysis, it was clear that MDV-transformed cell lines showed a distinct reduction in hybridization signals, especially when compared with the retrovirus-transformed lymphocyte cell lines HP45 and AVOL-1 (Fig. 2a). Several recent studies have highlighted the potential multiple roles of miR-155 in functions ranging from innate immune responses to oncogenicity (Garzon & Croce, 2008). The molecular mechanisms that drive down the expression of miR-155 in MDV-transformed cell lines are not known. However, some of its functions on targets such as Pu.1 could be rescued by MDV-miR-M4, a highly expressed MDV-1-encoded functional orthologue of miR-155 (Zhao *et al.*, 2009). Although the regulatory expression dynamics of miR-155 and MDV-miR-M4 are not understood fully, the existence of autoregulatory mechanisms of miR-155 expression mediated through a common set of targets cannot be ruled out. It is interesting that, in KSHV-infected primary effusion lymphoma cell lines, miR-155 was also found to be downregulated in favour of the KSHV-encoded miR-K12-11 homologue (Skalsky *et al.*, 2007).

The data from this study have enabled us to characterize the miRNome of MDV-transformed tumours. Although this has provided valuable insights into the expression profiles of miRNAs in these cell lines, the major challenge will be in the identification of the putative targets of the differentially expressed miRNAs in these cells. Although bioinformatic predictions of miRNA targets are valuable, the development of systems for functional characterization of miRNA targets is important to understand the pathways of oncogenesis. The tightly regulated, Dox-inducible miRNA expression system of the differentially expressed miRNAs that we developed in HEK293T cells will be valuable in identifying the putative functional targets of these miRNAs. Demonstration of the expression of mature miRNAs in a Dox-dependent manner clearly showed the proper processing of these miRNAs in this system. For functional validation of the system, we analysed the putative targeting of miR-155 on one of the validated target proteins, Pu.1 (Zhao *et al.*, 2009). The tightly regulated expression of miR-155 and the specific silencing of the relative luciferase levels with reporter assays with wild-type 3′ UTR reporter constructs (Fig. 3) provide a platform for functional analysis of the putative targets of differentially expressed miRNAs.

In summary, the data presented here demonstrate that miRNA expression profiling using microarrays is a powerful approach for analysing the relative levels of several miRNAs simultaneously. This study, the first of its kind in MDV-transformed cell lines, demonstrates that, in addition to the overexpression of several MDV-encoded miRNAs, downregulation of some of the host-encoded miRNAs is also a hallmark of MDV transformation. Determination of the miRNA profile is a first step towards identification of the regulatory networks of gene expression in these cell types.

ACKNOWLEDGEMENTS

The authors thank Dr Georg Bornkamm for providing the pRTS-1 plasmid used in the inducible expression systems. This work was carried out as part of a BBSRC grant awarded to V. N. We thank all members of the Avian Oncogenic Virus group for assistance and for inspiring discussions connected with this work, and Mick Gill for help with the digital imaging.

REFERENCES

Akiyama, Y. & Kato, S. (1974). Two cell lines from lymphomas of Marek's disease. *Biken J* **17**, 105–116.

Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P. & Bartel, D. P. (2008). The impact of microRNAs on protein output. *Nature* **455**, 64–71.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* **57**, 289–300.

Bornkamm, G. W., Berens, C., Kuklik-Roos, C., Bechet, J. M., Laux, G., Bachl, J., Korndoerfer, M., Schlee, M., Holzel, M. & other authors (2005). Stringent doxycycline-dependent control of gene activities using an episomal one-vector system. *Nucleic Acids Res* **33**, e137.

Burnside, J. & Morgan, R. W. (2007). Genomics and Marek's disease virus. *Cytogenet Genome Res* **117**, 376–387.

Burnside, J., Bernberg, E., Anderson, A., Lu, C., Meyers, B. C., Green, P. J., Jain, N., Isaacs, G. & Morgan, R. W. (2006). Marek's disease virus encodes microRNAs that map to *meq* and the latency-associated transcript. *J Virol* **80**, 8778–8786.

Burnside, J., Ouyang, M., Anderson, A., Bernberg, E., Lu, C., Meyers, B. C., Green, P. J., Markis, M., Isaacs, G. K. & other authors (2008). Deep sequencing of chicken microRNAs. *BMC Genomics* **9**, 185.

Cai, X., Lu, S., Zhang, Z., Gonzalez, C. M., Damania, B. & Cullen, B. R. (2005). Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc Natl Acad Sci U S A* **102**, 5570–5575.

Calin, G. A. & Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nat Rev Cancer* **6**, 857–866.

Calin, G. A. & Croce, C. M. (2007). Investigation of microRNA alterations in leukemias and lymphomas. *Methods Enzymol* **427**, 193–213.

Calnek, B. W. (1986). Marek's disease: a model for herpesvirus oncology. *Crit Rev Microbiol* **12**, 293–320.

Chan, M. M., Chen, C. L., Ager, L. L. & Cooper, M. D. (1988). Identification of the avian homologues of mammalian CD4 and CD8 antigens. *J Immunol* **140**, 2133–2138.

Cosmopoulos, K., Pegtel, M., Hawkins, J., Moffett, H., Novina, C., Middeldorp, J. & Thorley-Lawson, D. A. (2008). Comprehensive profiling of EBV microRNAs in nasopharyngeal carcinoma. *J Virol* **83**, 2357–2367.

Cullen, B. R. (2006). Viruses and microRNAs. *Nat Genet* **38** (Suppl.), S25–S30.

Cullen, B. R. (2009). Viral and cellular messenger RNA targets of viral microRNAs. *Nature* **457**, 421–425.

Fabbri, M., Garzon, R., Andreeff, M., Kantarjian, H. M., Garcia-Manero, G. & Calin, G. A. (2008). MicroRNAs and noncoding RNAs in hematological malignancies: molecular, clinical and therapeutic implications. *Leukemia* **22**, 1095–1105.

Garcia, P. & Frampton, J. (2008). Hematopoietic lineage commitment: miRNAs add specificity to a widely expressed transcription factor. *Dev Cell* **14**, 815–816.

Garzon, R. & Croce, C. M. (2008). MicroRNAs in normal and malignant hematopoiesis. *Curr Opin Hematol* **15**, 352–358.

Ghosh, Z., Mallick, B. & Chakrabarti, J. (2008). Cellular versus viral microRNAs in host–virus interaction. *Nucleic Acids Res* **37**, 1035–1048.

Gottwein, E. & Cullen, B. R. (2008). Viral and cellular microRNAs as determinants of viral pathogenesis and immunity. *Cell Host Microbe* **3**, 375–387.

Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**, D154–D158.

Javier, R. T. & Butel, J. S. (2008). The history of tumor virology. *Cancer Res* **68**, 7693–7706.

Johnnidis, J. B., Harris, M. H., Wheeler, R. T., Stehling-Sun, S., Lam, M. H., Kirak, O., Brummelkamp, T. R., Fleming, M. D. & Camargo, F. D. (2008). Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. *Nature* **451**, 1125–1129.

Lambeth, L. S., Yao, Y., Smith, L. P., Zhao, Y. & Nair, V. (2009). MicroRNAs 221 and 222 target $p27^{Kip1}$ in Marek's disease virus-transformed tumour cell line MSB-1. *J Gen Virol* **90**, 1164–1171.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O. & other authors (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414.

Lawrie, C. H., Soneji, S., Marafioti, T., Cooper, C. D., Palazzo, S., Paterson, J. C., Cattan, H., Enver, T., Mager, R. & other authors (2007). MicroRNA expression distinguishes between germinal center B cell-like and activated B cell-like subtypes of diffuse large B cell lymphoma. *Int J Cancer* **121**, 1156–1161.

Lawrie, C. H., Saunders, N. J., Soneji, S., Palazzo, S., Dunlop, H. M., Cooper, C. D., Brown, P. J., Troussard, X., Mossafa, H. & other authors (2008). MicroRNA expression in lymphocyte development and malignancy. *Leukemia* **22**, 1440–1446.

Lee, Y. S. & Dutta, A. (2009). MicroRNAs in cancer. *Annu Rev Pathol* **4**, 199–227.

Lin, Y. C., Kuo, M. W., Yu, J., Kuo, H. H., Lin, R. J., Lo, W. L. & Yu, A. L. (2008). c-Myb is an evolutionary conserved miR-150 target and miR-150/c-Myb interaction is important for embryonic development. *Mol Biol Evol* **25**, 2189–2198.

Lowery, A. J., Miller, N., McNeill, R. E. & Kerin, M. J. (2008). MicroRNAs as prognostic indicators and therapeutic targets: potential effect on breast cancer management. *Clin Cancer Res* **14**, 360–365.

Lu, J., Guo, S., Ebert, B. L., Zhang, H., Peng, X., Bosco, J., Pretz, J., Schlanger, R., Wang, J. Y. & other authors (2008). MicroRNA-mediated control of cell fate in megakaryocyte–erythrocyte progenitors. *Dev Cell* **14**, 843–853.

Martinez, I., Gardiner, A. S., Board, K. F., Monzon, F. A., Edwards, R. P. & Khan, S. A. (2008). Human papillomavirus type 16 reduces the expression of microRNA-218 in cervical carcinoma cells. *Oncogene* **27**, 2575–2582.

Merkerova, M., Belickova, M. & Bruchova, H. (2008). Differential expression of microRNAs in hematopoietic cell lineages. *Eur J Haematol* **81**, 304–310.

Morgan, R., Anderson, A., Bernberg, E., Kamboj, S., Huang, E., Lagasse, G., Isaacs, G., Parcells, M., Meyers, B. C. & other authors (2008). Sequence conservation and differential expression of Marek's disease virus microRNAs. *J Virol* **82**, 12213–12220.

Nazerian, K. (1987). An updated list of avian cell lines and transplantable tumours. *Avian Pathol* **16**, 527–544.

Ozen, M., Creighton, C. J., Ozdemir, M. & Ittmann, M. (2008). Widespread deregulation of microRNA expression in human prostate cancer. *Oncogene* **27**, 1788–1793.

Petherbridge, L., Brown, A. C., Baigent, S. J., Howes, K., Sacco, M. A., Osterrieder, N. & Nair, V. K. (2004). Oncogenicity of virulent Marek's disease virus cloned as bacterial artificial chromosomes. *J Virol* **78**, 13376–13380.

Pfeffer, S., Zavolan, M., Grasser, F. A., Chien, M., Russo, J. J., Ju, J., John, B., Enright, A. J., Marks, D. & other authors (2004). Identification of virus-encoded microRNAs. *Science* **304**, 734–736.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F. A., van Dyk, L. F., Ho, C. K., Shuman, S. & other authors (2005). Identification of microRNAs of the herpesvirus family. *Nat Methods* **2**, 269–276.

Pratt, Z. L., Kuzembayeva, M., Sengupta, S. & Sugden, B. (2009). The microRNAs of Epstein–Barr virus are expressed at dramatically differing levels among cell lines. *Virology* **386**, 387–397.

Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., Zepeniuk, M., Benjamin, H., Shabes, N., Tabak, S. & other authors (2008). MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol* **26**, 462–469.

Ruike, Y., Ichimura, A., Tsuchiya, S., Shimizu, K., Kunimoto, R., Okuno, Y. & Tsujimoto, G. (2008). Global correlation analysis for

micro-RNA and mRNA expression profiles in human cell lines. *J Hum Genet* **53**, 515–523.

Schotte, D., Chau, J. C., Sylvester, G., Liu, G., Chen, C., van der Velden, V. H., Broekhuis, M. J., Peters, T. C., Pieters, R. & Boer, M. L. (2008). Identification of new microRNA genes and aberrant microRNA profiles in childhood acute lymphoblastic leukemia. *Leukemia* **23**, 313–322.

Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N., Palmer, A. E. & Tsien, R. Y. (2004). Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nat Biotechnol* **22**, 1567–1572.

Skalsky, R. L., Samols, M. A., Plaisance, K. B., Boss, I. W., Riva, A., Lopez, M. C., Baker, H. V. & Renne, R. (2007). Kaposi's sarcoma-associated herpesvirus encodes an ortholog of miR-155. *J Virol* **81**, 12836–12845.

Smyth, G. K. (2005). limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pp. 397–420. Edited by R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry & S. Dudoit. New York: Springer.

Sullivan, C. S. & Grundhoff, A. (2007). Identification of viral microRNAs. *Methods Enzymol* **427**, 3–23.

Wiemer, E. A. (2007). The role of microRNAs in cancer: no small matter. *Eur J Cancer* **43**, 1529–1544.

Xiao, C., Calado, D. P., Galler, G., Thai, T. H., Patterson, H. C., Wang, J., Rajewsky, N., Bender, T. P. & Rajewsky, K. (2007). MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb. *Cell* **131**, 146–159.

Xu, H., Yao, Y., Zhao, Y., Smith, L. P., Baigent, S. J. & Nair, V. (2008). Analysis of the expression profiles of Marek's disease virus-encoded microRNAs by real-time quantitative PCR. *J Virol Methods* **149**, 201–208.

Yao, Y., Zhao, Y., Xu, H., Smith, L. P., Lawrie, C. H., Sewer, A., Zavolan, M. & Nair, V. (2007). Marek's disease virus type 2 (MDV-2)-encoded microRNAs show no sequence conservation with those encoded by MDV-1. *J Virol* **81**, 7164–7170.

Yao, Y., Zhao, Y., Xu, H., Smith, L. P., Lawrie, C. H., Watson, M. & Nair, V. (2008). MicroRNA profile of Marek's disease virus-transformed T-cell line MSB-1: predominance of virus-encoded microRNAs. *J Virol* **82**, 4007–4015.

Yeung, M. L., Yassunaga, J., Bennasser, Y., Dusetti, N., Harris, D., Ahmad, N., Matsuoka, M. & Jeang, K. T. (2008). Roles for microRNAs, miR-93 and miR-130b, and tumor protein 53-induced nuclear protein 1 tumor suppressor in cell growth dysregulation by human T-cell lymphotrophic virus 1. *Cancer Res* **68**, 8976–8985.

Zhao, Y., Yao, Y., Xu, H., Lambeth, L., Smith, L. P., Kgosana, L., Wang, X. & Nair, V. (2009). A functional MicroRNA-155 ortholog encoded by the oncogenic Marek's disease virus. *J Virol* **83**, 489–492.

Zhou, B., Wang, S., Mayr, C., Bartel, D. P. & Lodish, H. F. (2007). miR-150, a microRNA expressed in mature B and T cells, blocks early B cell development when expressed prematurely. *Proc Natl Acad Sci U S A* **104**, 7080–7085.

JOURNAL OF VIROLOGY, July 2009, p. 6969–6973 Vol. 83, No. 13
0022-538X/09/$08.00+0 doi:10.1128/JVI.00322-09

Novel MicroRNAs (miRNAs) Encoded by Herpesvirus of Turkeys: Evidence of miRNA Evolution by Duplication▽

Yongxiu Yao, Yuguang Zhao, Lorraine P. Smith, Michael Watson, and Venugopal Nair*

Division of Microbiology, Institute for Animal Health, Compton, Berkshire RG20 7NN, United Kingdom

Received 13 February 2009/Accepted 22 April 2009

Herpesviruses account for 134 out of the 140 virus-encoded microRNAs (miRNAs) known today. Here we report the identification of 11 novel miRNAs encoded by herpesvirus of turkey (HVT), a virus used as a live vaccine in poultry against the highly oncogenic Marek's disease virus type 1. Ten of these miRNAs were clustered together within the repeat long region of the viral genome, demonstrating some degree of positional conservation with other mardiviruses. Close sequence and phylogenetic relationships of some miRNAs in this cluster indicate evolution by duplication. HVT miRNAs represent the first example of virus-encoded miRNAs that show evolution by duplication.

MicroRNAs (miRNAs) are increasingly recognized as major regulators of gene expression in several multicellular organisms and viruses. *Herpesviridae*, a large family of viruses associated with a number of diseases including cancer in humans and animals, account for most of the virus-encoded miRNAs known today (7, 8). Considering the distinct biological requirements of herpesviruses, such as long latency periods that require the avoidance of the host immune responses, it is perhaps not very surprising that herpesviruses make extensive use of this highly effective regulatory mechanism of gene expression (7, 9, 21).

Marek's disease (MD) is a highly contagious rapid-onset T-cell lymphoma of poultry caused by MD virus type 1 (MDV-1), an alphaherpesvirus of the genus *Mardivirus* (11), which also includes MDV-2 and the herpesvirus of turkeys (HVT). Originally isolated from domestic turkeys in the late 1960s (12, 24), HVT is widely used as a live vaccine against MD because of its antigenic relatedness to MDV-1 (1). HVT is estimated to have separated from the MDV-1/MDV-2 lineage only about 38 million years ago (19). This relatedness is further evident from the overall similarities in the genome structures and sequences of these viruses (1, 14). We and others have previously reported the characteristics of several miRNAs encoded by MDV-1 and MDV-2 (5, 6, 25, 26). In the present study, we extended these investigations to identify HVT-encoded miRNAs. For this, we size selected the small RNA (~19- to 24-nucleotide [nt]) population from chicken embryo fibroblast (CEF) cultures infected with HVT by using procedures described previously (25). Sequence analysis of ~480 clones using vector-specific primers identified a total of 1,346 high-quality reads containing small RNA sequences with both the 5′ and 3′ adapters used in the cloning. BLAST homology searches (2) of the sequences against the HVT genome (AF291866) identified a total of 406 sequences (30%) representing 11 candidate miRNAs. The complex steps involved in miRNA biogenesis, including the criteria for strand selection for incorporation in the RNA-induced silencing complex, have been described previously (4, 13). Although it is the miRNA strand that is predominantly incorporated into the RNA-induced silencing complex, the non-miRNA strand could also be processed less efficiently than miRNAs (23, 25, 26). In the present study, mature forms representing both strands of the duplex were demonstrated either by cloning or by Northern blotting for 8 of the 11 HVT miRNAs. The miRNAs derived from the 5′ and 3′ arms of the stem-loop precursors were designated with a -5p or -3p suffix, respectively (Fig. 1A).

Examination of the genomic locations of the 11 candidate miRNAs showed that 10 of these miRNAs were clustered together in the same orientation in a 2.1-kb region (positions 120864 to 122977 in the TR_L/IR_L region) in the HVT genome, overlapping with two putative open reading frames, HVT074 and HVT075 (Fig. 1B). These included miRNA-6 and miRNA-7 embedded in HVT074 and miRNA-10 in HVT075, respectively, in the same orientations. The only HVT miRNA outside this cluster, miRNA-11, was located in the U_L region in the same orientation to the coding region of tegument protein UL21. The clustering of the 10 miRNAs in the same orientation in the 2.1-kb region of the TR_L/IR_L suggests that these miRNAs could be derived from a single transcript. However, the cloning frequencies of the miRNAs from this cluster showed variations, with miRNA-7-3p being the most abundant, with 207 hits representing more than half of the HVT miRNA population. In contrast, miRNAs such as miRNA-1-3p, miRNA-8-3p, and miRNA-11 were represented only once (Fig. 1A). Such differences in the abundance of the miRNAs have also been observed with MDV-1 and MDV-2 miRNAs (25, 26) and are thought to be due to differences in their processing efficiency or stability.

For validation of the authenticity of the candidate miRNAs, we used the standard criterion of miRNA annotation (3). First, we examined the potential precursor RNA hairpin structures of each of the 11 putative HVT miRNA candidates by using the 60- to 80-nt surrounding HVT genome sequence by MFOLD calculation (27). Secondary structures drawn using RNADRAW software showed that the miRNA precursors with average lengths of approximately 70 nt each were able to

* Corresponding author. Mailing address: Division of Microbiology, Institute for Animal Health, Compton, Berkshire RG20 7NN, United Kingdom. Phone: 441635 577356. Fax: 441635 577263. E-mail: venu.gopal@bbsrc.ac.uk.

▽ Published ahead of print on 29 April 2009.

A **Sequence and genomic position of HVT miRNAs**

Name	Sequences 5' to 3'	Length (nt)	Method of detection/number of hits	Genomic Position
HVT-miR-1-5p	AUUUUUCCUAAAGCGUGCCCGGGU	24	Northern blot	120864:120887
HVT-miR-1-3p	CCGGUCCGUUGUAGGAAUUAGG	22	1	120903:120924
HVT-miR-2-5p	UCUCAUGGACCCCCGAGUUGU(G)	21-22	17	121408:121429
HVT-miR-2-3p	(CA)GCUAGGCCGUCCAUGAGGUC(GUA)	21-23	16	121442:121466
HVT-miR-3	UCCUUCUCGCGGAUAGCGUAGAUG(U)	24-25	5	121553:121577
HVT-miR-4-5p	(UU)UCUUAUGGACCGCCGAGGUG(UGUC)	21-25	9	121726:121751
HVT-miR-4-3p	(CA)GCUAGGCCGUCCAUGAGGUC(GUA)	21-23	16	121762:121786
HVT-miR-5	(U)UUUCUCCGGUGAGCGUAGC(UGCC)	20-24	30	121872:121895
HVT-miR-6-5p	(UC)AUUCAGCGGGCAAUGUAGAC(UGU)	22-25	23	122098:122122
HVT-miR-6-3p	(A)GCUACAUUGCCCGCUGGGUUU(C)	22	4	122135:122157
HVT-miR-7-5p	UCGGAACAGGCGUAUUUUUUGC	22	2	122226:122247
HVT-miR-7-3p	(CA)AAGAUACGCAUGGGCUG(AGGGA)	18-24	207	122264:122287
HVT-miR-8-5p	UUCAUGUCGGGGGUGUUUGUG	21	Northern blot	122510:122530
HVT-miR-8-3p	UAAAUAGCCCCGAUUUGACAG	21	1	122547:122567
HVT-miR-9-5p	UUCACGUCGGGGGUGUUUGUG	21	Northern blot	122674:122694
HVT-miR-9-3p	UAAACAGCCCCGAUUUGACA(GGU)	20-23	71	122711:122733
HVT-miR-10-5p	(A)UCAAACCGGGGGGGAUUUCAGU	22-23	3	122916:122938
HVT-miR-10-3p	CUGAGAUACGCCCGAUUUGACGAA	24	Northern blot	122954:122977
HVT-miR-11	GGGAUGAAAUGGGCUUUACCAU	22	1	42536:42557

B

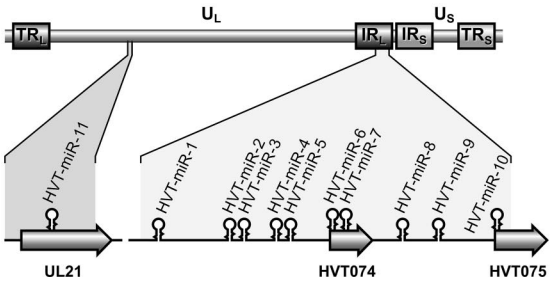


FIG. 1. Cloning of HVT miRNAs. (A) Sequences of cloned candidate miRNA species, with nucleotide positions based on the published sequence (GenBank accession number AF291866). The methods of detection of the miRNAs or their cloning frequencies in the library are indicated. Sequence variations surrounding the recovered HVT miRNAs are indicated by parentheses. (B) The schematic diagram showing the positions of the miRNAs (stem-loops) in the HVT genome. The internal and terminal repeat long (TR_L/IR_L) and short (TR_S/IR_S) regions flanking the unique long (U_L) and short (U_S) regions of the genome are shown. Genomic positions and orientations of HVT open reading frames UL21, HVT074, and HVT075 are shown.

form characteristic hairpin structures (Fig. 2A), supporting the view that the cloned sequences represent novel HVT-encoded miRNAs. We next examined the expression of the mature HVT miRNAs by Northern blot analysis. For this, the total RNA isolated from HVT-infected CEFs was hybridized with individual miRNA probes. RNA from uninfected CEFs was included as a negative control. All of the cloned sequences are detectable in HVT-infected CEFs by Northern blotting, except for miR-1-3p, which appeared only once in the library (Fig. 2B). The other strand of miR-1-3p precursor is detectable, suggesting that miR-1-5p could be the miRNA strand and that miR-1-3p could serve as the non-miRNA strand. No miRNAs were detected with RNA extracted from the uninfected CEFs, although weak signals were observed with the miR-10-5p probe, which also gave a strong band between the precursor and mature forms in the HVT-infected cells. This band, as well as the mature band, also appears in the uninfected CEFs with much lower intensity. The reasons behind the detection of these signals in uninfected cells are not clear. Although BLAST searches did not reveal any similar sequences in the chicken genome, this cannot be ruled out, especially because of the incomplete annotation of the chicken genome sequence. Northern blotting detected the miRNA as well as the non-miRNA strands in several cases, although the latter generally showed lower expression levels based on the signal intensities (Fig. 2B). This is also reflected by the lower ratio of the mature non-miRNA strand to the pre-miRNA.

The discovery of 11 novel HVT-encoded miRNAs reported here together with the previously identified 14 in MDV-1 (5, 6, 20, 26) and 17 in MDV-2 (25) makes the genus *Mardivirus* a major source of virus-encoded miRNAs. Closer examination of some of the HVT miRNA sequences showed striking similarities, suggesting their generation by duplication. Gene duplication has long been recognized as a major route for evolution of genes including miRNAs in several species (10, 16–18). Duplications of miRNAs are generally observed in large miRNA clusters, suggesting that the expansion of the miRNA clusters is a major mode of miRNA evolution (10). In order to obtain evidence of miRNA duplication in the HVT miRNA cluster, we carried out multiple sequence alignment of the precursors of these miRNAs. As shown in Fig. 3A, the sequence of the precursors of miR-2 and miR-4 showed high sequence homology (95.3%) with identical sequences in the loop and the mature -3p regions, while the -5p mature miRNA region showed three substitutions, including the one in the seed region. Similarly, the precursors of miR-8 and miR-9 are highly homologous with only a single nucleotide difference in both -5p and -3p mature miRNA sequences as well as in the loop regions. The close sequence homology of the HVT miRNAs is also evident from the phylogenetic analysis which showed the branching of the closely related miRNAs (Fig. 3B). Because they are antisense regulators, alteration to the miRNA sequences of duplicated miRNAs can have a major impact on their targeting capabilities and capacities for acquiring novel functions, particularly for changes in the seed regions. The duplicated HVT miRNAs miR-8 and miR-9 had

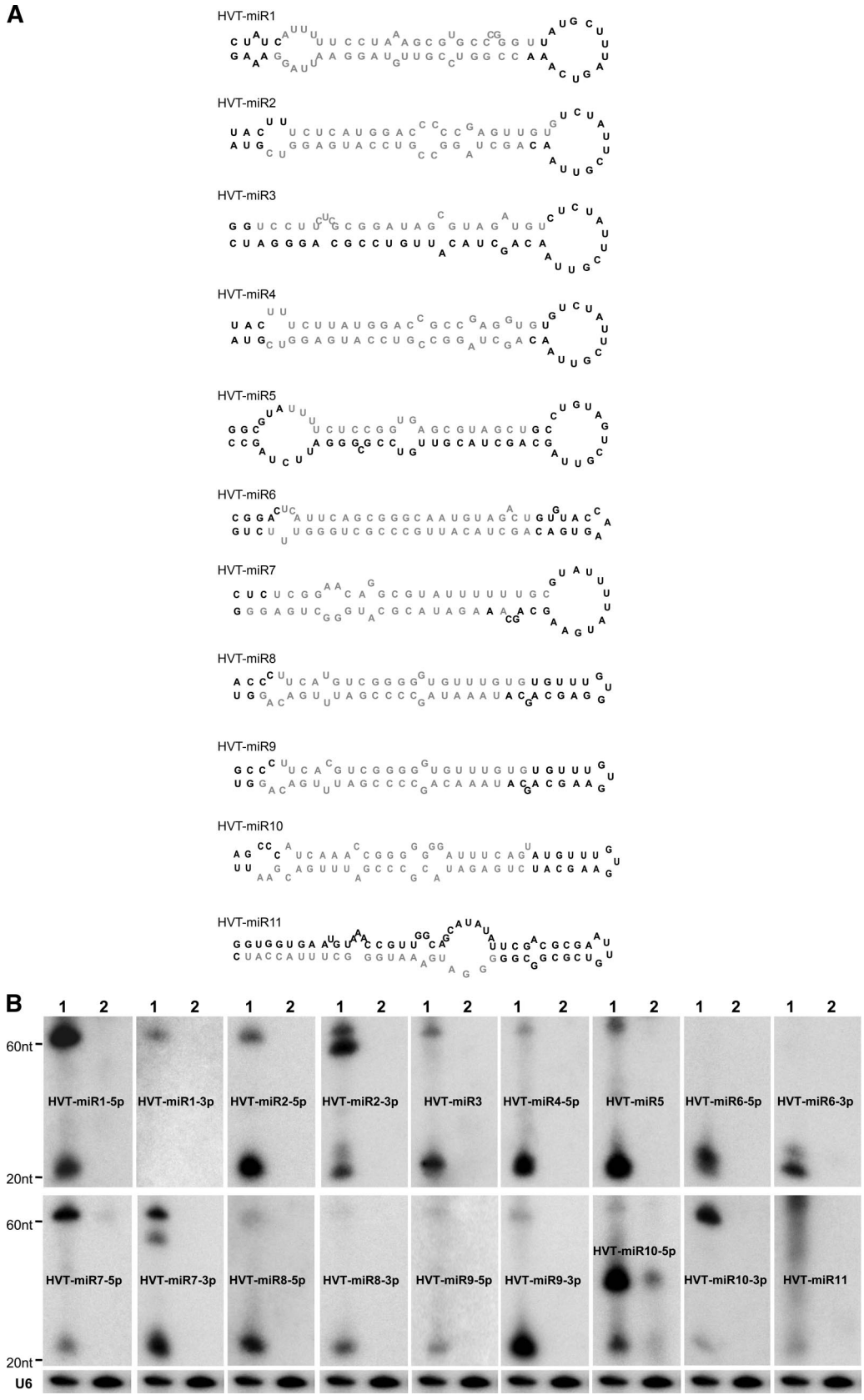


FIG. 2. Identification of cloned HVT miRNAs. (A) Secondary structures of HVT pre-miRNAs predicted using the MFOLD algorithm. The mature miRNA strands are indicated in light gray. (B) Northern blot analysis demonstrating the expression of HVT miRNAs. Total RNAs from HVT-infected CEFs (lanes 1) and uninfected CEFs (lanes 2) were separated on a 15% denaturing polyacrylamide gel and probed with [γ-^{32}P]ATP-radiolabeled antisense oligonucleotides to the indicated miRNAs. Size markers indicate the positions of the pre-miRNA and the mature miRNA. The cellular U6 small nuclear RNA served as the loading control. A representative blot of this set is shown.

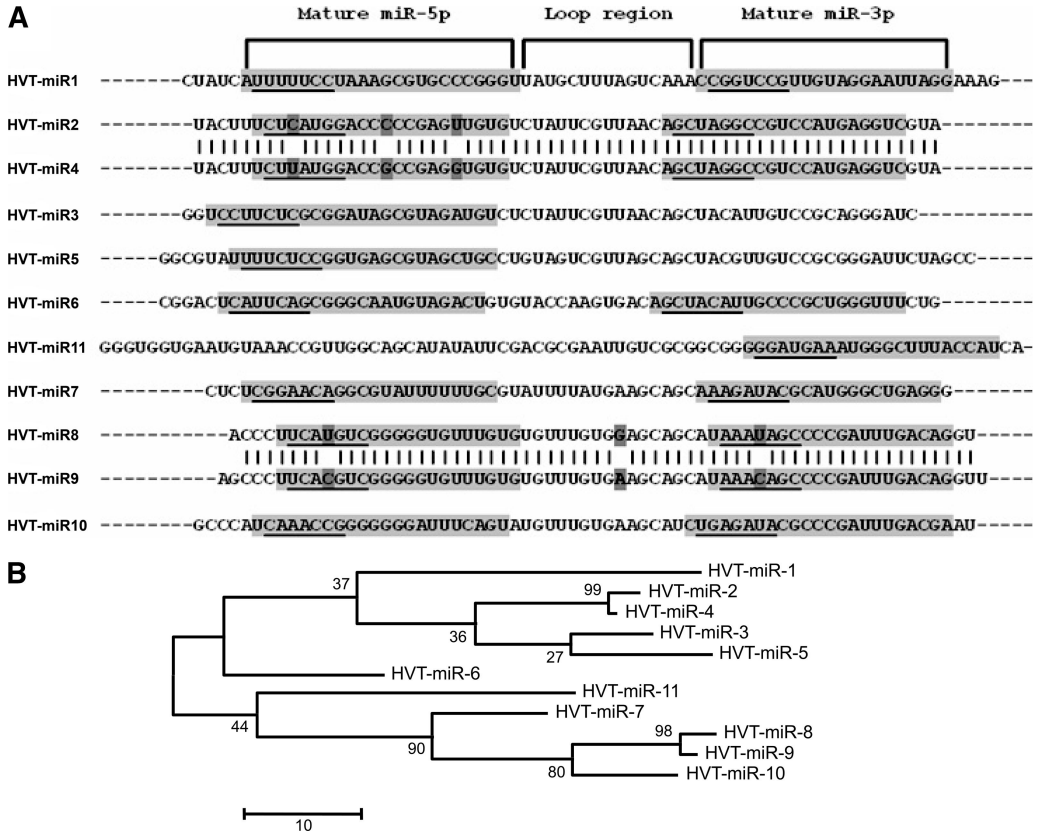


FIG. 3. Phylogeny of HVT miRNAs. (A) Sequence alignment of HVT pre-miRNAs showing the positions of the mature -5p and -3p miRNAs (shaded areas) and the loop region. The seed regions of the mature miRNAs are underlined. (B) Phylogenetic tree (with the bootstrap values of the branches) showing the evolutionary relationships of the HVT pre-miRNAs by maximum parsimony analysis using the MEGA 4.0 package (15).

identical sequences in both -5p and -3p mature miRNA sequences, except for a single point mutation in the seed sequence (Fig. 3A). Since any transcript is just 1 nt away from being an miRNA target, such point mutations in the seed sequence could result in the loss or gain of new targets for these miRNAs, allowing "neofunctionalization" (22) of these duplicated miRNAs. Similarly, the three nucleotide substitutions including the one in the seed region of the -5p mature sequences of miR-2 and miR-4 will also have the potential to affect the expression of new targets. Although further work is needed to examine the targets of these novel miRNAs, HVT-encoded miRNAs represent the first clear example of evolution of miRNAs by duplication among viruses.

We thank Nick Knowles, Institute for Animal Health, Pirbright, for assistance with the phylogenetic analysis of the miRNAs.

This work was partly funded by the BBSRC, United Kingdom.

REFERENCES

1. **Afonso, C. L., E. R. Tulman, Z. Lu, L. Zsak, D. L. Rock, and G. F. Kutish.** 2001. The genome of turkey herpesvirus. J. Virol. **75:**971–978.
2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
3. **Ambros, V., B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl.** 2003. A uniform system for microRNA annotation. RNA **9:**277–279.
4. **Bartel, D. P.** 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell **116:**281–297.
5. **Burnside, J., E. Bernberg, A. Anderson, C. Lu, B. C. Meyers, P. J. Green, N. Jain, G. Isaacs, and R. W. Morgan.** 2006. Marek's disease virus encodes microRNAs that map to *meq* and the latency-associated transcript. J. Virol. **80:**8778–8786.
6. **Burnside, J., M. Ouyang, A. Anderson, E. Bernberg, C. Lu, B. C. Meyers, P. J. Green, M. Markis, G. Isaacs, E. Huang, and R. W. Morgan.** 2008. Deep sequencing of chicken microRNAs. BMC Genomics **9:**185.
7. **Cullen, B. R.** 2009. Viral and cellular messenger RNA targets of viral microRNAs. Nature **457:**421–425.
8. **Gottwein, E., and B. R. Cullen.** 2008. Viral and cellular microRNAs as determinants of viral pathogenesis and immunity. Cell Host Microbe **3:**375–387.
9. **Grey, F., L. Hook, and J. Nelson.** 2008. The functions of herpesvirus-encoded microRNAs. Med. Microbiol. Immunol. **197:**261–267.
10. **Hertel, J., M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I. L. Hofacker, and P. F. Stadler.** 2006. The expansion of the metazoan microRNA repertoire. BMC Genomics **7:**25.
11. **ICTV.** 2006. 00.031.1.03 *Mardivirus. In* C. Búchen-Osmond (ed.), ICTVdB—the universal virus database, version 4. Columbia University, New York, NY.
12. **Kawamura, H., D. J. King, Jr., and D. P. Anderson.** 1969. A herpesvirus isolated from kidney cell culture of normal turkeys. Avian Dis. **13:**853–863.
13. **Kim, V. N., J. Han, and M. C. Siomi.** 2009. Biogenesis of small RNAs in animals. Nat. Rev. Mol. Cell Biol. **10:**126–139.
14. **Kingham, B. F., V. Zelnik, J. Kopacek, V. Majerciak, E. Ney, and C. J. Schmidt.** 2001. The genome of herpesvirus of turkeys: comparative analysis with Marek's disease viruses. J. Gen. Virol. **82:**1123–1135.
15. **Kumar, S., M. Nei, J. Dudley, and K. Tamura.** 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. **9:**299–306.
16. **Liu, N., K. Okamura, D. M. Tyler, M. D. Phillips, W. J. Chung, and E. C. Lai.** 2008. The evolution and functional diversification of animal microRNA genes. Cell Res. **18:**985–996.
17. **Lu, J., Y. Fu, S. Kumar, Y. Shen, K. Zeng, A. Xu, R. Carthew, and C. I. Wu.** 2008. Adaptive evolution of newly emerged micro-RNA genes in Drosophila. Mol. Biol. Evol. **25:**929–938.
18. **Lu, J., Y. Shen, Q. Wu, S. Kumar, B. He, S. Shi, R. W. Carthew, S. M. Wang, and C. I. Wu.** 2008. The birth and death of microRNA genes in Drosophila. Nat. Genet. **40:**351–355.

19. **McGeoch, D. J., F. J. Rixon, and A. J. Davison.** 2006. Topics in herpesvirus genomics and evolution. Virus Res. **117:**90–104.
20. **Morgan, R., A. Anderson, E. Bernberg, S. Kamboj, E. Huang, G. Lagasse, G. Isaacs, M. Parcells, B. C. Meyers, P. J. Green, and J. Burnside.** 2008. Sequence conservation and differential expression of Marek's disease virus microRNAs. J. Virol. **82:**12213–12220.
21. **Murphy, E., J. Vanicek, H. Robins, T. Shenk, and A. J. Levine.** 2008. Suppression of immediate-early viral gene expression by herpesvirus-coded microRNAs: implications for latency. Proc. Natl. Acad. Sci. USA **105:**5453–5458.
22. **Rastogi, S., and D. A. Liberles.** 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol. **5:**28.
23. **Seo, G. J., L. H. Fink, B. O'Hara, W. J. Atwood, and C. S. Sullivan.** 2008. Evolutionarily conserved function of a viral microRNA. J. Virol. **82:**9823–9828.
24. **Witter, R. L., K. Nazerian, H. G. Purchase, and G. H. Burgoyne.** 1970. Isolation from turkeys of a cell-associated herpesvirus antigenically related to Marek's disease virus. Am. J. Vet. Res. **31:**525–538.
25. **Yao, Y., Y. Zhao, H. Xu, L. P. Smith, C. H. Lawrie, A. Sewer, M. Zavolan, and V. Nair.** 2007. Marek's disease virus type 2 (MDV-2)-encoded microRNAs show no sequence conservation with those encoded by MDV-1. J. Virol. **81:**7164–7170.
26. **Yao, Y., Y. Zhao, H. Xu, L. P. Smith, C. H. Lawrie, M. Watson, and V. Nair.** 2008. MicroRNA profile of Marek's disease virus-transformed T-cell line MSB-1: predominance of virus-encoded microRNAs. J. Virol. **82:**4007–4015.
27. **Zuker, M.** 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31:**3406–3415.

Journal of General Virology (2012), **93**, 1530–1536 **DOI** 10.1099/vir.0.040634-0

Short Communication

Novel microRNAs encoded by duck enteritis virus

Yongxiu Yao,[1] Lorraine P. Smith,[1] Lawrence Petherbridge,[1] Mick Watson[2] and Venugopal Nair[1]

[1]Viral Oncogenesis Group, Institute for Animal Health, Compton, Berkshire RG20 7NN, UK

[2]Ark-Genomics, The Roslin Institute, R(D)SVS, University of Edinburgh, Division of Genetics and Genomics, Easter Bush, Midlothian EH25 9RG, UK

Correspondence
Venugopal Nair
venugopal.nair@iah.ac.uk

Received 16 December 2011
Accepted 30 March 2012

Duck enteritis virus (DEV) is an important herpesvirus pathogen associated with acute, highly contagious lethal disease in waterfowls. Using a deep sequencing approach on RNA from infected chicken embryo fibroblast cultures, we identified several novel DEV-encoded micro (mi)RNAs. Unlike most mardivirus-encoded miRNAs, DEV-encoded miRNAs mapped mostly to the unique long region of the genome. The precursors of DEV miR-D18 and miR-D19 overlapped with each other, suggesting similarities to miRNA-offset RNAs, although only the DEV-miR-D18-3p was functional in reporter assays. Identification of these novel miRNAs will add to the growing list of virus-encoded miRNAs enabling the exploration of their roles in pathogenesis.

MicroRNAs (miRNAs) are increasingly recognized as major regulators of gene expression in many organisms, including viruses. Herpesviruses, belonging to α-, β- and γ-families, account for the majority of the currently known virus-encoded miRNAs (Cullen, 2009; Gottwein & Cullen, 2008). Duck enteritis virus (DEV), also referred to as the duck plague virus, is a highly contagious α-herpesvirus causing an acute disease with high mortality in waterfowl. A recent study has shown that DEV has a genome structure similar to those of iltoviruses (Zhong *et al.*, 2009). We and others have previously reported the identification of miRNAs from a number of avian herpesviruses including 14 miRNAs from Marek's disease virus (MDV)-1 (Burnside *et al.*, 2006; Yao *et al.*, 2008), 18 from MDV-2 (Waidner *et al.*, 2009; Yao *et al.*, 2007), 17 from herpesvirus of turkeys (HVT) (Waidner *et al.*, 2009; Yao *et al.*, 2009) and seven from infectious laryngeotracheitis virus (ILTV) (Rachamadugu *et al.*, 2009; Waidner *et al.*, 2009). The functional targets of most of these miRNAs are not known. Nevertheless, some of these miRNAs have been shown to play major roles in virus pathogenesis (Burnside & Morgan, 2011; Yao *et al.*, 2009). Most virus-encoded miRNAs have shown little sequence conservation with those encoded by other viruses, although their locations in the viral genomes did reveal some levels of conservation (Waidner *et al.*, 2009; Yao *et al.*, 2007, 2009). To examine whether DEV did encode miRNAs, we carried out deep-sequence analysis on small RNA extracted from chicken embryo fibroblast (CEF) (miRVana miRNA isolation kit; Ambion) heavily infected with anatid herpesvirus-1 strain 568 (kindly provided by Professor Kaleta, University of Giessen, Germany) on the Illumina GAIIx platform by GATC Biotech. Out of the 1 345 371 sequence reads with high base quality scores, 34 644 aligned to the DEV genome sequence (GenBank accession no. EU082088). Of these, a total of 19 251 that perfectly matched with the DEV sequence represented 46 candidate miRNAs. Using the RNA folding program Mfold (Zuker, 2003), the 60–80 nt region surrounding each of the DEV miRNAs could be folded into primary miRNA hairpin structures, with each mature miRNA forming one arm of the stem. For further confirmation of the potential DEV-encoded miRNAs, array-based miRNA profiling was performed on DEV-infected CEF and uninfected CEF using miRNA microarrays designed to cover all chicken miRNAs in miRBase v.17 (http://www.mirbase.org/) and the candidate DEV-encoded miRNAs identified by deep-sequencing analysis. The miRNA microarray expression assay and the statistical analysis of the microarray data were performed by LC Sciences. Probes for the chicken miRNAs (a total of 542 miRNAs; miRBase release 17) and 72 custom probes of candidate DEV miRNAs were printed on chips in six replicates and hybridized with Cy5-labelled small RNAs isolated from either DEV-infected CEF or uninfected CEF. In addition, the control probes were included on each chip for quality controls of chip production, sample labelling and assay conditions. Data were analysed using ANOVA and *t*-tests. Normalization of expression was performed using a cyclic LOWESS method. miRNAs showing differential expression at the *P*-value <0.01 were selected for clustering analysis, which was performed using a hierarchical method based on average linkage and a Euclidean distance metric (Eisen *et al.*, 1998). In agreement with the sequencing data, all the miRNAs undetectable by microarray are those with very low frequency particularly with only one read from the deep sequencing. The miRNAs displaying differential expression (*P*-values <0.01) by cluster analysis were analysed further and sequentially

Supplementary tables are available with the online version of this paper.

named from DEV-miR-D1 to DEV-miR-D24 according to their genomic locations. Mature miRNA species showed sequence lengths of 19–24 nt, most of them 22 nt long. The numbers of individual reads of each of these 24 miRNAs varied greatly. While some of the miRNAs such as DEV-miR-D8 were highly abundant with 3028 reads, others such as DEV-miR-D2 was detected only as a single copy but they are still differentially expressed as detected by miRNA microarray (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36258). The functional relevance of miRNAs detected at very low levels by deep sequencing has been questioned (Cullen, 2011). However, such low copy miRNAs have also been detected in other viruses using other assays such as Northern blotting or RT-PCR (Umbach & Cullen, 2010; Waidner *et al.*, 2009). The features of the 24 DEV-encoded miRNAs, including the nine miRNAs where both strands of the mature miRNA duplex could be demonstrated, are shown in Table S1 (available in JGV Online). The secondary structure of those miRNAs is shown in Fig. 1(a). Sequence length variability was noted at both the 5′- and 3′-ends of the miRNAs as indicated by parentheses in Table S1, generating families of isomiRs for some of the miRNAs possibly arising from the imprecision in precursor processing (Morin *et al.*, 2008). The 5′-end of miRNAs, especially the 2–7 nt 'seed sequence', is critical for target mRNA binding and translational inhibition function (Bartel, 2009). Therefore, the sequence differences at the 5′-end of some DEV miRNAs provide the potential to target multiple mRNA targets (Gottwein & Cullen, 2008). We also examined whether any of the DEV miRNAs showed sequence homology with existing miRNAs in the miRBase (http://www.mirbase.org/). For this, each of the DEV miRNA sequences were searched against those in the miRBase database (Release 18) using a regular expression search implemented in Perl (http://www.perl.org) that looked for seed as well as full-length sequence matches in other species. Although this search did not identify any full-length homologies with other miRNAs, some of the DEV miRNAs showed seed sequence homology with a number of miRNAs in the database (Table S2). Our previous study demonstrating the ability of MDV-miR-M4 and gga-miR-155 to act as functional homologues despite only sharing their seed regions (Zhao *et al.*, 2009), indicated the potential of the DEV miRNAs also to act as functional homologues of miRNAs sharing the seed sequence. For gaining further insights into the DEV miRNA functions, we examined the predicted targets of the three highly expressed DEV-encoded miRNAs using the miRNA target prediction program Targetscan 5.2 Custom (http://www.targetscan.org) against chicken 3′-UTR. Most of the target sites were conserved across multiple species (data not shown) thus increasing their likelihood to be the real targets. The numbers of predicted targets of each of three miRNAs varied greatly. While there were over 200 predicted targets for the most highly abundant DEV-miR-D8-3p, the second highly expressed DEV-miR-D17-5p has only one predicted target. There was nine predicted target genes for DEV-miR-D18-3p (Table S3). Due to the large number of the predicted target genes for DEV-miR-D8-3p, only those with $\geqslant 2$ target sites are shown. A number of these target genes have been reported to be involved in other virus infections. These include the GRIA4 (glutamate receptor, ionotrophic, AMPA 4) gene differentially expressed in Newcastle disease virus infection (Lan *et al.*, 2010), KPNA3 (karyopherin α3) involved in the nuclear import of the influenza virus (Carter, 2009), HMGB1 (high mobility group box 1) protein involved in the release of both RNA (Barqasho *et al.*, 2010; Chen *et al.*, 2008; Chu & Ng, 2003; Jung *et al.*, 2011; Kamau *et al.*, 2009; Wang *et al.*, 2006) and DNA viruses (Borde *et al.*, 2011) and NCAM1 (neural cell adhesion molecule 1) that interacts with the spike protein of coronaviruses (Gao *et al.*, 2010). Although further studies are needed to identify the precise functions of DEV miRNAs, it is likely that these would involve the modulation of the functions of some of these molecules.

In addition to the differentially expressed DEV miRNAs, cluster analysis also indicated striking differences in cellular miRNA expression between DEV-infected and uninfected CEF (Fig. 2). All significantly expressed miRNAs (P-value <0.01) with mean intensity values of more than 500 (arbitrary units) were selected for inclusion in the heatmaps. Expression levels of 45 chicken miRNAs were altered by DEV infection, of these, 26 miRNAs were upregulated and 19 were downregulated. For the upregulated miRNAs, only two showed a more than twofold increase: gga-miR-203 and gga-miR-1607. Seven of the downregulated miRNAs showed a more than twofold decrease in their expression: miR-1759, miR-1767, miR-466, miR-146b, miR-1690*, miR-1796, and miR-1560*. Interestingly, miR-146b previously reported to be associated with immune-related signal pathways in mammals also downregulated in avian influenza virus-infected lungs and tracheae in chicken (Lindsay, 2008; Wang *et al.*, 2009). The exact function of the differentially expressed miRNAs in the DEV infection remains to be determined.

Unlike most mardivirus-encoded miRNAs that are located at the repeat region (Fig. 1b), the majority of the DEV miRNAs were encoded within the unique long region as six clusters from both the coding and non-coding regions of the 15 809 bp viral genome (Fig. 1c). The miRNAs miR-D22 and miR-D23 were encoded from the coding region of ICP4 in an antisense orientation, enabling them to function in an siRNA-like fashion (Barth *et al.*, 2008; Seo *et al.*, 2008, 2009; Sullivan *et al.*, 2005; Tang *et al.*, 2008, 2009; Waidner *et al.*, 2011), as reported in MDV-2 and ILTV (Waidner *et al.*, 2009; Yao *et al.*, 2007). Discovery of novel DEV-encoded miRNAs together with the previously identified miRNAs in MDV-1 (Burnside *et al.*, 2006; Morgan *et al.*, 2008; Waidner *et al.*, 2009; Yao *et al.*, 2008), MDV-2 (Waidner *et al.*, 2009; Yao *et al.*, 2007), HVT (Waidner *et al.*, 2009; Yao *et al.*, 2009) and ILTV (Waidner *et al.*, 2011), make avian α-herpesviruses a rich source for viral miRNAs.

Closer examination of the genomic locations of the DEV miRNAs revealed that the precursor sequences of two of

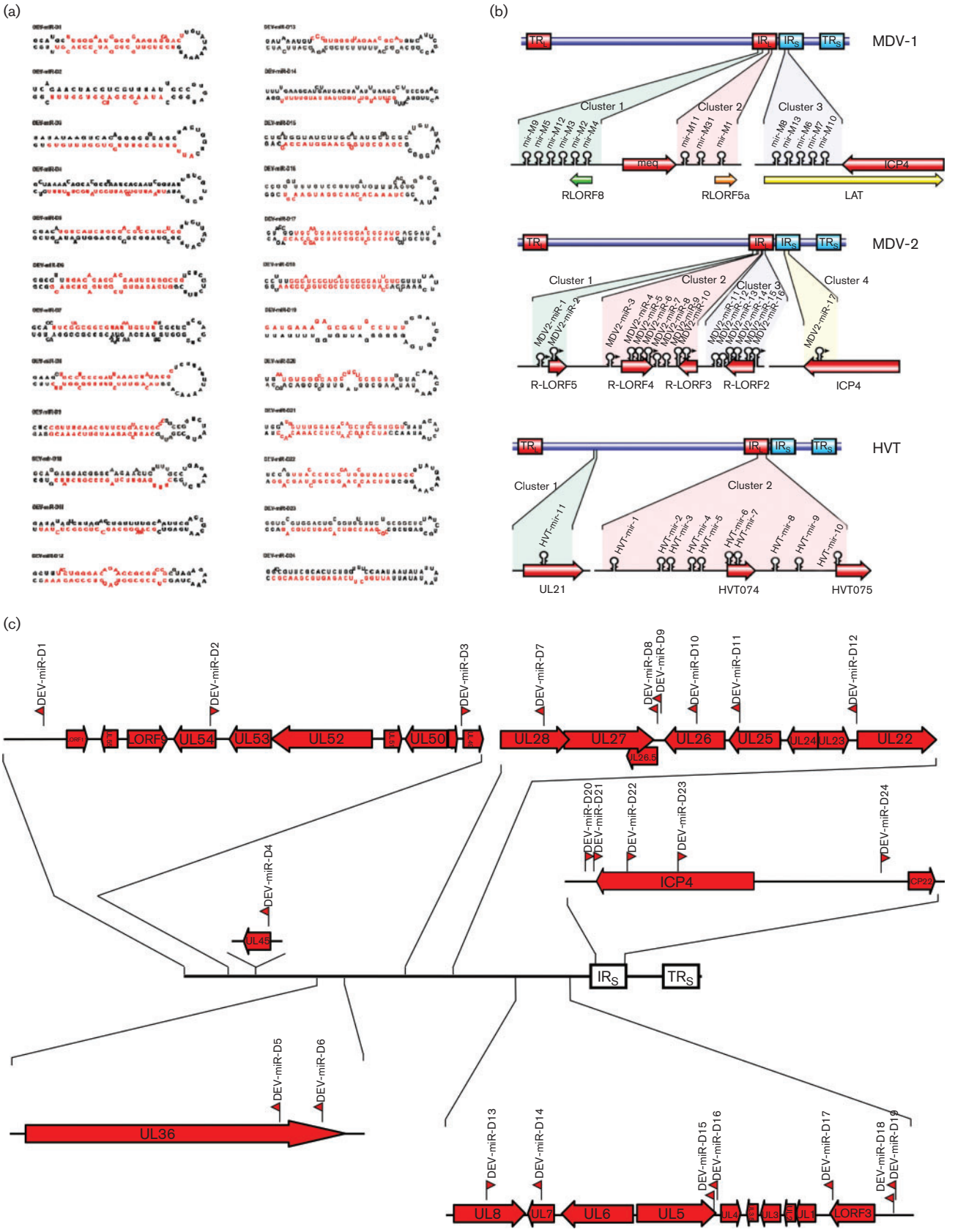

(a)
(b)
MDV-1
TRL
IRL
IRS
TRS
Cluster 1
Cluster 2
Cluster 3
mir-M9
mir-M5
mir-M12
mir-M3
mir-M2
mir-M4
mir-M11
mir-M31
mir-M1
mir-M8
mir-M13
mir-M6
mir-M7
mir-M10
meq
ICP4
RLORF8
RLORF5a
LAT
MDV-2
Cluster 4
MDV2-miR-1
MDV2-miR-2
MDV2-miR-3
MDV2-miR-4
MDV2-miR-5
MDV2-miR-6
MDV2-miR-7
MDV2-miR-8
MDV2-miR-9
MDV2-miR-10
MDV2-miR-11
MDV2-miR-12
MDV2-miR-13
MDV2-miR-14
MDV2-miR-15
MDV2-miR-16
MDV2-miR-17
R-LORF5
R-LORF4
R-LORF3
R-LORF2
ICP4
HVT
HVT-mir-11
HVT-mir-1
HVT-mir-2
HVT-mir-3
HVT-mir-4
HVT-mir-5
HVT-mir-6
HVT-mir-7
HVT-mir-8
HVT-mir-9
HVT-mir-10
UL21
HVT074
HVT075
(c)
DEV-miR-D1
DEV-miR-D2
DEV-miR-D3
DEV-miR-D7
DEV-miR-D8
DEV-miR-D9
DEV-miR-D10
DEV-miR-D11
DEV-miR-D12
ORF1
LORF6
UL54
UL53
UL52
UL51
UL50
UL49
UL28
UL27
UL26.5
UL26
UL25
UL24
UL23
UL22
DEV-miR-D20
DEV-miR-D21
DEV-miR-D22
DEV-miR-D23
DEV-miR-D24
ICP4
CP22
DEV-miR-D4
UL45
IRS
TRS
DEV-miR-D5
DEV-miR-D6
UL36
DEV-miR-D13
DEV-miR-D14
DEV-miR-D15
DEV-miR-D16
DEV-miR-D17
DEV-miR-D18
DEV-miR-D19
UL8
UL7
UL6
UL5
UL4
UL3.5
UL3
UL2
UL1
LORF3

Fig. 1. Structure and genomic locations of DEV miRNAs (a) Mfold prediction of the secondary structures of DEV pre-miRNAs with the mature miRNA strands are indicated in red; (b) genome structure of mardiviruses demonstrating the restriction of miRNA clusters mostly to the terminal and internal short and long (TRS, IRS, TRL and IRL) repeat regions; (c) DEV genome structure showing the internal (IR) and terminal (TR) repeat regions, with the genomic positions and direction of transcription of miRNAs (indicated by flags). The orientations of each of the ORFs in relation to the miRNA loci are also indicated.

the miRNAs, miR-D18 and -D19 do overlap (Fig. 3a). The mature sequence of the two miRNAs follows the same pattern as miRNA offset RNAs (moRNAs), a recently discovered class of small RNAs (Shi *et al.*, 2009) closely related to miRNAs (Fig. 3b). moRNAs are derived from sequences located immediately adjacent to the mature miRNA and miRNA* strands in the pri-miRNA precursor and have been recovered at low levels in several studies using deep sequencing (Babiarz *et al.*, 2008; Jurak *et al.*, 2010; Ruby *et al.*, 2007; Shi *et al.*, 2009; Umbach *et al.*, 2010). The stem–loop of miR-D19 does not seem stable based on some of the set criteria (Han *et al.*, 2006). However, the high level expression of miR-D19 in infected cells as evident from the number of reads by deep sequencing (Table S1), signal intensity of 9158 arbitrary units in microarray analysis (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36258), detection in Northern blot analysis (Fig. 3c) and high expression levels in infected cells (Fig. 3e) indicate that DEV-miR-D19 is a genuine miRNA. In order to examine the functionality of these two miRNAs, we carried out reporter assays using *Renilla* luciferase-based reporter plasmids (Zhao *et al.*, 2009), bearing four tandem repeats of artificial target sites with a perfect match to each miRNA, inserted into the 3′-UTR. The reporter plasmid was co-transfected into DF-1 cells along with the plasmid expressing miR-D18 and -D19. The luciferase expression was assayed 36 h post-transfection using the Dual-Glo Luciferase Assay System (Promega). While miR-D18 was able to inhibit luciferase reporter expression 90 % relative to a negative control reporter plasmid, miR-D19 showed no inhibitory effect (Fig. 3d), demonstrating that only miR-D18 is potentially functional. In order to validate the expression of both miRNAs in this assay, we determined the expression levels in transfected cells using the TaqMan miRNA Assay System (Applied Biosystems) together with the DEV-infected CEF and normal DF-1 controls. The result showed that miR-D18 and -D19 are equally expressed although at a low level compared with the infected cells (Fig. 3e). Despite our efforts to demonstrate functionality to DEV-miR-D19, the expression pattern of DEV-miR-D18 and D19 identified in this report represent the first example of the overlapping property of adjacent miRNAs.

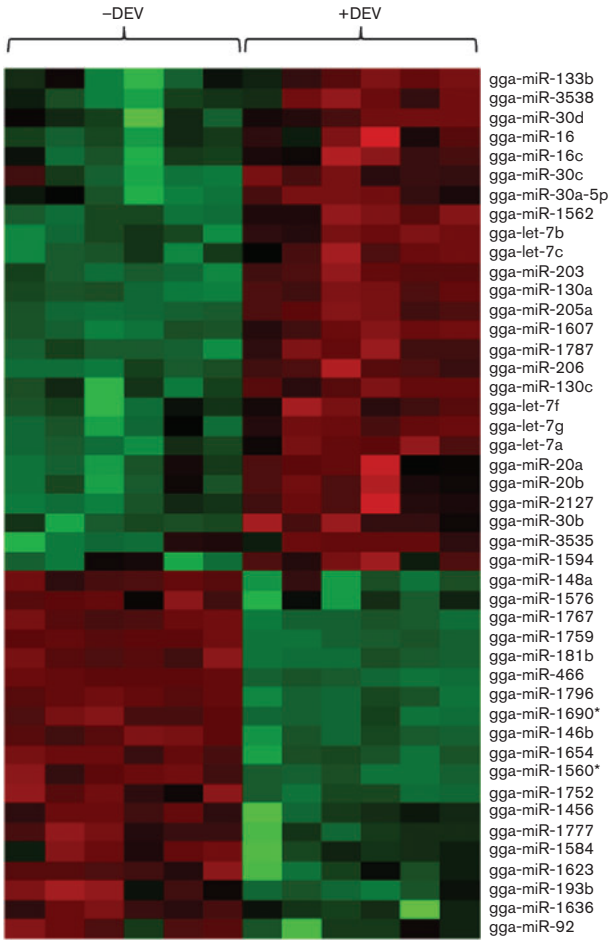


Fig. 2. DEV infection alters cellular miRNA expression profile, as determined using microarrays. Expression profiles of miRNAs that were significantly up or downregulated in DEV-infected CEF compared with uninfected CEF are shown. Each miRNA was replicated six times on each chip. Red and green indicate upregulation and downregulation of expression, respectively.

Identification of novel miRNAs reported here provides the opportunity for further studies to examine the role of these novel miRNAs in DEV biology and pathogenesis. A recent report comparing the genome sequence of the virulent wild-type DEV strain 2085 and the vaccine strain of DEV has detected differences in 54 of 78 predicted ORFs (Wang *et al.*, 2011), but not in any of the miRNAs reported in this study. However, these observations are not entirely surprising as pathogenic and vaccine strains of MDV also did not show difference with respect to miRNAs (Xu *et al.*, 2008). A more direct approach to examine the functions of these miRNAs in the disease pathogenesis will be to use viruses with deletions in specific miRNAs in disease models (Zhao *et al.*, 2011). A recent report on the development of infectious bacterial artificial chromosome clones of DEV (Wang & Osterrieder, 2011) would enable the generation of miRNA mutant viruses using reverse genetics approaches.

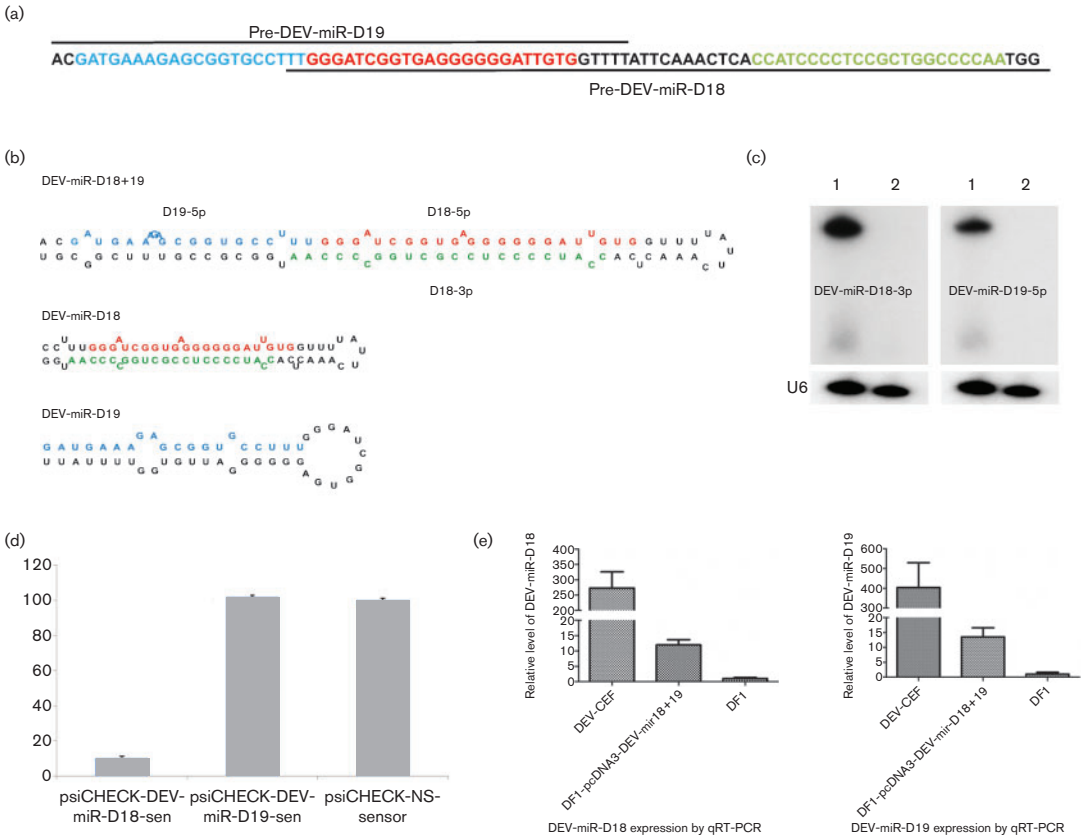


Fig. 3. DEV-miR-D19 is a genuine miRNA. (a and b) The sequence of the individual and combined pre-miRNA-D18 and pre-miR-D19 showing the overlapping property of the two miRNAs. The sequence of mature miR-D18-3p is indicated in green and the miR-D18-5p star strand shown in red. The sequence of mature miR-D19-5p is indicated in teal. (c) Northern blot analysis demonstrating the expression of mature and pre-miRNA of DEV-miR-D18 and -D19. Total RNAs from DEV-infected CEFs (lanes 1) and uninfected CEFs (lanes 2) were separated on a 15 % denaturing polyacrylamide gel and probed with [γ-^{32}P]ATP-radiolabelled antisense oligonucleotides to the indicated miRNAs. The cellular U6 small nuclear RNA served as the loading control. A representative blot of this set is shown. (d) Inhibitory activity of miR-D18-3p and miR-D19-5p revealed in transfected DF-1 cells using luciferase-based indicator plasmids bearing four perfectly complementary target sites. For each sample, values from four replicates representative of at least two independent experiments were used in the analysis. Values are displayed relative to a control indicator bearing an unrelated sequence, which was set at 100 %. The histogram shows the relative levels of *Renilla* luciferase in DF-1 cells co-transfected with reporter vectors and DEV-miR-D18 and -D19 expression construct. (e) miRNA expression levels determined by qRT-PCR. Relative expression of miR-D18-3p and miR-D19-5p measured in RNA extracted from DEV-infected CEF, miR-D18 and miR-D19 expression plasmid transfected DF-1 and untransfected DF-1. Results represent the mean of triplicate assays with error bars showing SEM.

Acknowledgements

This work was supported by Biotechnology and Biological Sciences Research Council, UK. Authors thank Dr Andrew Dalby for assisting in sequence matching of miRNAs of various species.

References

Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P. & Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other microprocessor-independent, dicer-dependent small RNAs. *Genes Dev* **22**, 2773–2785.

Barqasho, B., Nowak, P., Abdurahman, S., Walther-Jallow, L. & Sönnerborg, A. (2010). Implications of the release of high-mobility group box 1 protein from dying cells during human immunodeficiency virus type 1 infection *in vitro*. *J Gen Virol* **91**, 1800–1809.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233.

Barth, S., Pfuhl, T., Mamiani, A., Ehses, C., Roemer, K., Kremmer, E., Jäker, C., Höck, J., Meister, G. & Grässer, F. A. (2008). Epstein–Barr virus-encoded microRNA miR-BART2 down-regulates the viral DNA polymerase BALF5. *Nucleic Acids Res* **36**, 666–675.

Borde, C., Barnay-Verdier, S., Gaillard, C., Hocini, H., Maréchal, V. & Gozlan, J. (2011). Stepwise release of biologically active HMGB1 during HSV-2 infection. *PLoS ONE* **6**, e16145.

Burnside, J. & Morgan, R. (2011). Emerging roles of chicken and viral microRNAs in avian disease. *BMC Proc* **5** (Suppl. 4), S2.

Burnside, J., Bernberg, E., Anderson, A., Lu, C., Meyers, B. C., Green, P. J., Jain, N., Isaacs, G. & Morgan, R. W. (2006). Marek's disease virus encodes microRNAs that map to meq and the latency-associated transcript. *J Virol* **80**, 8778–8786.

Carter, C. J. (2009). Schizophrenia susceptibility genes directly implicated in the life cycles of pathogens: cytomegalovirus, influenza, herpes simplex, rubella, and toxoplasma gondii. *Schizophr Bull* **35**, 1163–1182.

Chen, L. C., Yeh, T. M., Wu, H. N., Lin, Y. Y. & Shyu, H. W. (2008). Dengue virus infection induces passive release of high mobility group box 1 protein by epithelial cells. *J Infect* **56**, 143–150.

Chu, J. J. & Ng, M. L. (2003). The mechanism of cell death during West Nile virus infection is dependent on initial infectious dose. *J Gen Virol* **84**, 3305–3314.

Cullen, B. R. (2009). Viral and cellular messenger RNA targets of viral microRNAs. *Nature* **457**, 421–425.

Cullen, B. R. (2011). Viruses and microRNAs: RISCy interactions with serious consequences. *Genes Dev* **25**, 1881–1894.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863–14868.

Gao, W., He, W., Zhao, K., Lu, H., Ren, W., Du, C., Chen, K., Lan, Y., Song, D. & Gao, F. (2010). Identification of NCAM that interacts with the PHE-CoV spike protein. *Virol J* **7**, 254.

Gottwein, E. & Cullen, B. R. (2008). Viral and cellular microRNAs as determinants of viral pathogenesis and immunity. *Cell Host Microbe* **3**, 375–387.

Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., Sohn, S. Y., Cho, Y., Zhang, B. T. & Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**, 887–901.

Jung, J. H., Park, J. H., Jee, M. H., Keum, S. J., Cho, M. S., Yoon, S. K. & Jang, S. K. (2011). Hepatitis C virus infection is blocked by HMGB1 released from virus-infected cells. *J Virol* **85**, 9359–9368.

Jurak, I., Kramer, M. F., Mellor, J. C., van Lint, A. L., Roth, F. P., Knipe, D. M. & Coen, D. M. (2010). Numerous conserved and divergent microRNAs expressed by herpes simplex viruses 1 and 2. *J Virol* **84**, 4659–4672.

Kamau, E., Takhampunya, R., Li, T., Kelly, E., Peachman, K. K., Lynch, J. A., Sun, P. & Palmer, D. R. (2009). Dengue virus infection promotes translocation of high mobility group box 1 protein from the nucleus to the cytosol in dendritic cells, upregulates cytokine production and modulates virus replication. *J Gen Virol* **90**, 1827–1835.

Lan, D., Tang, C., Li, M. & Yue, H. (2010). Screening and identification of differentially expressed genes from chickens infected with Newcastle disease virus by suppression subtractive hybridization. *Avian Pathol* **39**, 151–159.

Lindsay, M. A. (2008). microRNAs and the immune response. *Trends Immunol* **29**, 343–351.

Morgan, R., Anderson, A., Bernberg, E., Kamboj, S., Huang, E., Lagasse, G., Isaacs, G., Parcells, M., Meyers, B. C. & other authors (2008). Sequence conservation and differential expression of Marek's disease virus microRNAs. *J Virol* **82**, 12213–12220.

Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A. L., Zhao, Y., McDonald, H., Zeng, T. & other authors (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* **18**, 610–621.

Rachamadugu, R., Lee, J. Y., Wooming, A. & Kong, B. W. (2009). Identification and expression analysis of infectious laryngotracheitis virus encoding microRNAs. *Virus Genes* **39**, 301–308.

Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P. & Lai, E. C. (2007). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res* **17**, 1850–1864.

Seo, G. J., Fink, L. H., O'Hara, B., Atwood, W. J. & Sullivan, C. S. (2008). Evolutionarily conserved function of a viral microRNA. *J Virol* **82**, 9823–9828.

Seo, G. J., Chen, C. J. & Sullivan, C. S. (2009). Merkel cell polyomavirus encodes a microRNA with the ability to autoregulate viral gene expression. *Virology* **383**, 183–187.

Shi, W., Hendrix, D., Levine, M. & Haley, B. (2009). A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat Struct Mol Biol* **16**, 183–189.

Sullivan, C. S., Grundhoff, A. T., Tevethia, S., Pipas, J. M. & Ganem, D. (2005). SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature* **435**, 682–686.

Tang, S., Bertke, A. S., Patel, A., Wang, K., Cohen, J. I. & Krause, P. R. (2008). An acutely and latently expressed herpes simplex virus 2 viral microRNA inhibits expression of ICP34.5, a viral neurovirulence factor. *Proc Natl Acad Sci U S A* **105**, 10931–10936.

Tang, S., Patel, A. & Krause, P. R. (2009). Novel less-abundant viral microRNAs encoded by herpes simplex virus 2 latency-associated transcript and their roles in regulating ICP34.5 and ICP0 mRNAs. *J Virol* **83**, 1433–1442.

Umbach, J. L. & Cullen, B. R. (2010). In-depth analysis of Kaposi's sarcoma-associated herpesvirus microRNA expression provides insights into the mammalian microRNA-processing machinery. *J Virol* **84**, 695–703.

Umbach, J. L., Strelow, L. I., Wong, S. W. & Cullen, B. R. (2010). Analysis of rhesus rhadinovirus microRNAs expressed in virus-induced tumors from infected rhesus macaques. *Virology* **405**, 592–599.

Waidner, L. A., Morgan, R. W., Anderson, A. S., Bernberg, E. L., Kamboj, S., Garcia, M., Riblet, S. M., Ouyang, M., Isaacs, G. K. & other authors (2009). MicroRNAs of Gallid and Meleagrid herpesviruses show generally conserved genomic locations and are virus-specific. *Virology* **388**, 128–136.

Waidner, L. A., Burnside, J., Anderson, A. S., Bernberg, E. L., German, M. A., Meyers, B. C., Green, P. J. & Morgan, R. W. (2011). A microRNA of infectious laryngotracheitis virus can downregulate and direct cleavage of ICP4 mRNA. *Virology* **411**, 25–31.

Wang, J. & Osterrieder, N. (2011). Generation of an infectious clone of duck enteritis virus (DEV) and of a vectored DEV expressing hemagglutinin of H5N1 avian influenza virus. *Virus Res* **159**, 23–31.

Wang, H., Ward, M. F., Fan, X. G., Sama, A. E. & Li, W. (2006). Potential role of high mobility group box 1 in viral infectious diseases. *Viral Immunol* **19**, 3–9.

Wang, Y., Brahmakshatriya, V., Zhu, H., Lupiani, B., Reddy, S. M., Yoon, B. J., Gunaratne, P. H., Kim, J. H., Chen, R. & other authors (2009). Identification of differentially expressed miRNAs in chicken lung and trachea with avian influenza virus infection by a deep sequencing approach. *BMC Genomics* **10**, 512.

Wang, J., Höper, D., Beer, M. & Osterrieder, N. (2011). Complete genome sequence of virulent duck enteritis virus (DEV) strain 2085 and comparison with genome sequences of virulent and attenuated DEV strains. *Virus Res* **160**, 316–325.

Xu, H., Yao, Y., Zhao, Y., Smith, L. P., Baigent, S. J. & Nair, V. (2008). Analysis of the expression profiles of Marek's disease virus-encoded microRNAs by real-time quantitative PCR. *J Virol Methods* **149**, 201–208.

Yao, Y., Zhao, Y., Xu, H., Smith, L. P., Lawrie, C. H., Sewer, A., Zavolan, M. & Nair, V. (2007). Marek's disease virus type 2 (MDV-2)-encoded microRNAs show no sequence conservation with those encoded by MDV-1. *J Virol* **81**, 7164–7170.

Yao, Y., Zhao, Y., Xu, H., Smith, L. P., Lawrie, C. H., Watson, M. & Nair, V. (2008). MicroRNA profile of Marek's disease virus-transformed T-cell line MSB-1: predominance of virus-encoded microRNAs. *J Virol* **82**, 4007–4015.

Yao, Y., Zhao, Y., Smith, L. P., Watson, M. & Nair, V. (2009). Novel microRNAs (miRNAs) encoded by herpesvirus of turkeys: evidence of miRNA evolution by duplication. *J Virol* **83**, 6969–6973.

Zhao, Y., Yao, Y., Xu, H., Lambeth, L., Smith, L. P., Kgosana, L., Wang, X. & Nair, V. (2009). A functional microRNA-155 ortholog encoded by the oncogenic Marek's disease virus. *J Virol* **83**, 489–492.

Zhao, Y., Xu, H., Yao, Y., Smith, L. P., Kgosana, L., Green, J., Petherbridge, L., Baigent, S. J. & Nair, V. (2011). Critical role of the virus-encoded microRNA-155 ortholog in the induction of Marek's disease lymphomas. *PLoS Pathog* **7**, e1001305.

Zhong, Z., Chai, T., Duan, H., Miao, Z., Li, X., Yao, M., Yuan, W., Wang, W., Li, Q. & other authors (2009). REP-PCR tracking of the origin and spread of airborne *Staphylococcus aureus* in and around chicken house. *Indoor Air* **19**, 511–516.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406–3415.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 14 August 2013
doi: 10.3389/fgene.2013.00153

MicroRNA expression profiles in avian haemopoietic cells

*Yongxiu Yao[1], Jane Charlesworth[2], Venugopal Nair[1] and Mick Watson[2]**

[1] Avian Viral Diseases Programme, Compton Laboratory, The Pirbright Institute, Berkshire, UK
[2] ARK-Genomics, Royal (Dick) School of Veterinary Studies, The Roslin Institute, University of Edinburgh, Edinburgh, UK

Edited by:
Yingqun Huang, Yale University School of Medicine, USA

Reviewed by:
Mark T. McNally, Medical College of Wisconsin, USA
Ling-Ling Chen, Institute of Biochemistry and Cell Biology, China

***Correspondence:**
Mick Watson, ARK-Genomics, Royal (Dick) School of Veterinary Studies, The Roslin Institute, University of Edinburgh, Easter Bush, EH25 9RG, Edinburgh, UK
e-mail: mick.watson@roslin.ed.ac.uk

MicroRNAs (miRNAs) are small, abundant, non-coding RNAs that modulate gene expression by interfering with translation or stability of mRNA transcripts in a sequence-specific manner. A total of 734 precursor and 996 mature miRNAs have so far been identified in the chicken genome. A number of these miRNAs are expressed in a cell type-specific manner, and understanding their function requires detailed examination of their expression in different cell types. We carried out deep sequencing of small RNA populations isolated from stimulated or transformed avian haemopoietic cell lines to determine the changes in the expression profiles of these important regulatory molecules during these biological events. There were significant changes in the expression of a number of miRNAs, including miR-155, in chicken B cells stimulated with CD40 ligand. Similarly, avian leukosis virus (ALV)-transformed DT40 cells also showed changes in miRNA expression in relation to the naïve cells. Embryonic stem cell line BP25 demonstrated a distinct cluster of upregulated miRNAs, many of which were shown previously to be involved in embryonic stem cell development. Finally, chicken macrophage cell line HD11 showed changes in miRNA profiles, some of which are thought to be related to the transformation by v-*myc* transduced by the virus. This work represents the first publication of a catalog of microRNA expression in a range of important avian cells and provides insights into the potential roles of miRNAs in the hematopoietic lineages of cells in a model non-mammalian species.

Keywords: microRNA, B-cell, macrophages, DT40, HD11, IAH30

INTRODUCTION

MicroRNAs (miRNAs) are a large class of endogenous non-coding RNAs 21–23 nucleotides in length. Since they were first described nearly 20 years ago, there has been a steady increase in their discovery and the latest Release 20 of miRBase has 24,521 entries of miRNAs from various species, including 734 mature miRNAs from *Gallus gallus* (www.mirbase.org). Many of these miRNAs are expressed differentially during development in a cell-specific manner. A number of previous studies in mammals have demonstrated hematopoietic lineage-specific expression of miRNAs, suggesting key roles for these molecules in controlling hematopoietic machinery (Ramkissoon et al., 2006; Merkerova et al., 2008). Chickens are used as a model organism for a number of studies and the developing chicken embryo has been shown to be an excellent biological system to study the repertoire and dynamics of small regulatory RNAs (Glazov et al., 2008). The development of deep sequencing technologies and bioinformatics pipelines has greatly facilitated the discovery and quantification of expression levels of miRNAs in different cell types (Friedlander et al., 2008). We and others have previously reported the expression profiles of miRNAs in chicken T cells transformed by Marek's disease virus (MDV), and shown that the majority of the miRNAs expressed in these cell types are of viral origin (Burnside et al., 2008; Yao et al., 2009; Morgan and Burnside, 2011). Elevated expression of miRNAs such as gga-miR-155 has also been demonstrated in chicken hematopoietic cells transformed by reticuloendotheliosis virus (Bolisetty et al., 2009). However, studies examining the global expression of miRNAs in different haemopoietic cell lineages in chickens have not yet been carried out.

In this study we carried out deep sequencing of the miRNAs of six avian haemopoietic cell populations: **BP25**, a chick embryonic stem cell (cESC) line; **Bu1B**, naïve embryonic B lymphocytes; **StimB**, CD40L-induced B-cells; **DT40**, an avian-leukosis virus (ALV) transformed B-cell line; **HD11**, a chicken macrophage cell line; and **IAH30**, a turkey macrophage cell line. We have determined the miRNA expression profile of the cESC line, and we have compared the miRNA expression profiles of naïve B cells with B-cells after stimulation with the CD40 ligand (CD40L), to gain an understanding of the global changes in miRNA expression after signaling through the CD40 ligand interaction. We have also determined the miRNA profiles in the ALV-transformed B-cell line DT40 (Bachl et al., 2007) to identify the effects of the c-*myc*-induced transformation on miRNA expression. To further examine the effects of *myc*-induced transformation on other cell types of the haemopoietic lineage, we have determined the miRNA expression profiles of the chicken macrophage cell line HD11 (Beug et al., 1979). In addition, we also examined the miRNAs in the turkey macrophage cell line IAH30 (Lawson et al., 2001) to compare the miRNA profiles of chicken and turkey macrophages.

By analyzing the changes in expression of miRNA populations in the different cell types, our study provides insights into the cell type-specific miRNA signatures of avian haemopoietic cells. We believe that our data will be helpful in identifying targets and pathways associated with a number of phenotypic and functional characteristics of these lineages of avian haemopoietic cells.

MATERIALS AND METHODS

We made use of six cell lines for the characterization of the miRNAs: BP25 cESC line was propagated on irradiated STO feeder cells (ATCC collection) as previously described (Pain et al., 1996). Naïve B cell population was prepared from embryonic bursa of Fabricius (BF) or spleen collected from line 0 eggs at 18-day-old of embryonation. Briefly, BF was dissected from embryos and cell suspensions were separated by density gradient centrifugation on Ficoll-Paque under sterile conditions. B-cell preparations of more than 95% purity were obtained by magnetic cell sorting on MACS separation columns LS (Miltenyi Biotec, UK) using chicken B-cell (Bu-1)-specific monoclonal antibody AV20 (Rothwell et al., 1996) and anti-mouse microbeads as previously described (Kothlow et al., 2008). CD40L-induced *in vitro* B-cell proliferation was carried out as previously described using purified recombinant protein (Tregaskes et al., 2005; Kothlow et al., 2008), and cells were harvested 48 h after treatment with the ligand. DT40 (Buerstedde et al., 2002; Bachl et al., 2007), HD11 (Beug et al., 1979) and IAH30 (Lawson et al., 2001) cell lines were propagated as previously described.

RNA extraction for miRNA profiling was carried out as previously described (Yao et al., 2012) using miRVana miRNA isolation kit (Ambion, UK). Sequencing of the miRNAs was carried out on the Illumina GAIIx and 36 base-pair single-end sequencing. After sequencing, adaptor and primer/dimer sequences were removed using Cutadapt (http://code.google.com/p/cutadapt/). Using the Novoalign short read aligner (www.novocraft.com), we mapped the reads from all the individual cell lines, including the turkey macrophage cell line, to the known chicken mature miRNAs downloaded from miRBase (www.miRBase.org) version 19. Reads mapping to each miRNA were counted and used as input for downstream analyzes. To correct for differences in library size and sequencing depth, raw mapped read counts were scaled to reads per million mapped reads (Mortazavi et al., 2008). Changes in miRNA expression in CD40L-stimulated (StimB) cells compared to naïve B cells (Bu1B) were calculated as log2 ratios of normalized (RPM) counts. Similarly, changes in the expression of individual miRNAs in DT40 cells were also calculated in comparison to those of naïve B cells (BU1B). Normalized counts of miRNA levels were used to generate a heatmap in order to identify candidate miRNAs that are differentially expressed in different cell lines. The pearson correlation coefficient was used as a similarity measure in the heatmap cluster analyzis, and using the "average" agglomeration method. Validation of expression levels of gga-miR-21, gga-miR-26a, gga-miR142-3p, gga-miR-155 and gga-miR-223 was carried out by quantitative RT-PCR using procedures described (Yao et al., 2008).

RESULTS

RAW DATA

The raw data hava been submitted to the European Nucleotide Archive (ENA) under accession number ERP002558. Counts and normalized RPM values have been uploaded as Supplementary Material.

miRNA EXPRESSION IN B-LYMPHOCYTES AFTER CD40L STIMULATION

Naïve B cells are activated through a combination of signals from the antigen and through the binding of the CD40 ligand to CD40 which drives proliferation. Comparison of the naïve and CD40L-stimulated B cells revealed significant changes in the miRNA expression profiles (**Figure 1**). The miRNAs which showed significant increase upon CD40L-stimulation included gga-miR-21, gga-miR-155, gga-miR-146a, gga-miR-20b, gga-miR-106, gga-miR-222, and gga-miR-22. A number of miRNAs were also down-regulated after CD40L stimulation. This included gga-miR-26a which showed a 4.5-fold decrease in expression. Other down-regulated miRNAs included members of the miR-30 family of miRNAs (gga-miR-30c, gga-miR-30d, gga-miR-30a-5p) and the avian-specific gga-miR-1729 originally discovered in developing chick embryo.

miRNA EXPRESSION IN DT40 CELLS

Analyzis of the expression of miRNAs in DT40 cells demonstrated overexpression of a number of miRNAs compared to the naïve Bu1B positive cells (**Figure 2**). Among the upregulated miRNAs in DT40 cells, many miRNAs including gga-miR-18a and -18b, -222, -20b, -148a, -221, -106, -103, -101 and -21 also increased in CD40L-stimulated cells (**Figure 3A**). On the other hand, gga-miR-100 is highly upregulated in DT40 cells compared to naïve B cells or CD40L-stimulated cells (data not shown). Similarly, gga-miR-146a, upregulated in CD40L-stimulated B cells, was down-regulated in DT40 cells, providing further evidence for its role in the immune system. A number of miRNAs, including gga-miR-16, -30e, -30d, -30b, -30c, -26a, -147, -15b, and -29a, down-regulated in DT40 cells were also downregulated in CD40L-treated cells (**Figure 3B**). The most striking change was in the level of gga-miR-155. This miRNA was the most up-regulated miRNA in CD40L-stimulated B cells, but was downregulated in the DT40 cells.

Expression levels of miRNAs obtained from the deep sequencing data were validated by carrying out quantitative RT-PCR on gga-miR-21, gga-miR-26a, gga-miR-142-3p and gga-miR-155 using RNA extracted from different cell types. Differences in the expression profiles of the miRNAs determined from deep sequencing broadly agreed with the quantitative RT-PCR data from naïve Bu1B-positive, DT40 and stimulated B cells (**Figure 4A**). Similarly, the high expression levels of gga-miR-142-3p and gga-miR-223 in HD11 cells was also confirmed by quantitative RT-PCR (**Figure 4B**).

miRNA EXPRESSION IN AVIAN MACROPHAGE CELL LINES HD11 AND IAH30

Out of the 718,959 sequences in HD11 cells that mapped to the mature miRNAs in miRBase, the highest level of expression was

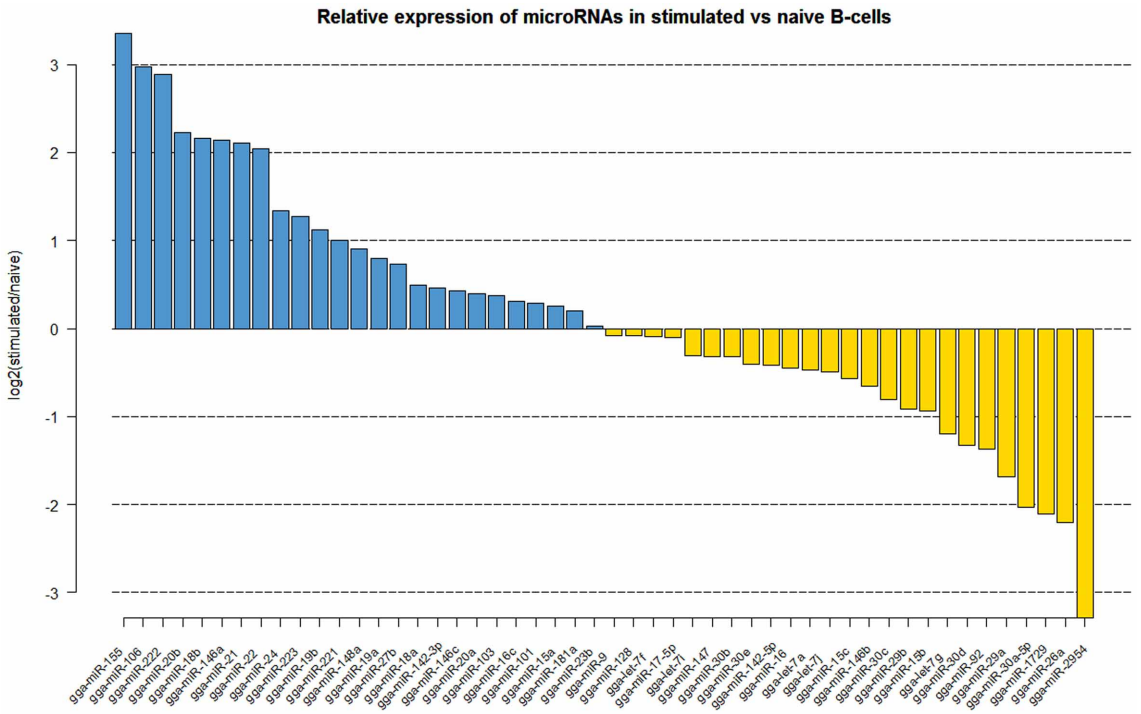


FIGURE 1 | Log2 fold (stimulated/naïve) change in expression (red or green bars indicating increased or decreased expression respectively) of the 50 most abundantly expressed miRNAs in CD40L-stimulated B cells (StimB) compared to the naïve B cells (BU1B).

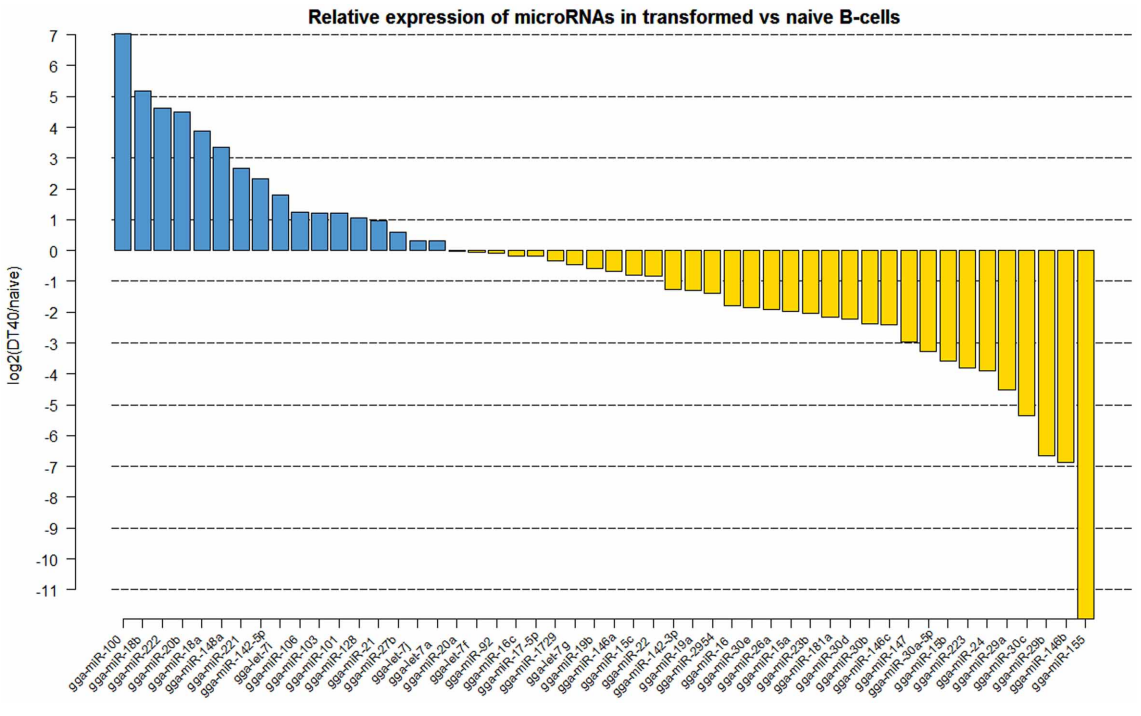


FIGURE 2 | Log2 fold (DT40/naïve) change in expression (red or green bars indicating increased or decreased expression respectively) of the 50 most abundantly expressed miRNAs in DT40 ALV-transformed B cells (DT40) compared to the naïve B cells (BU1B).

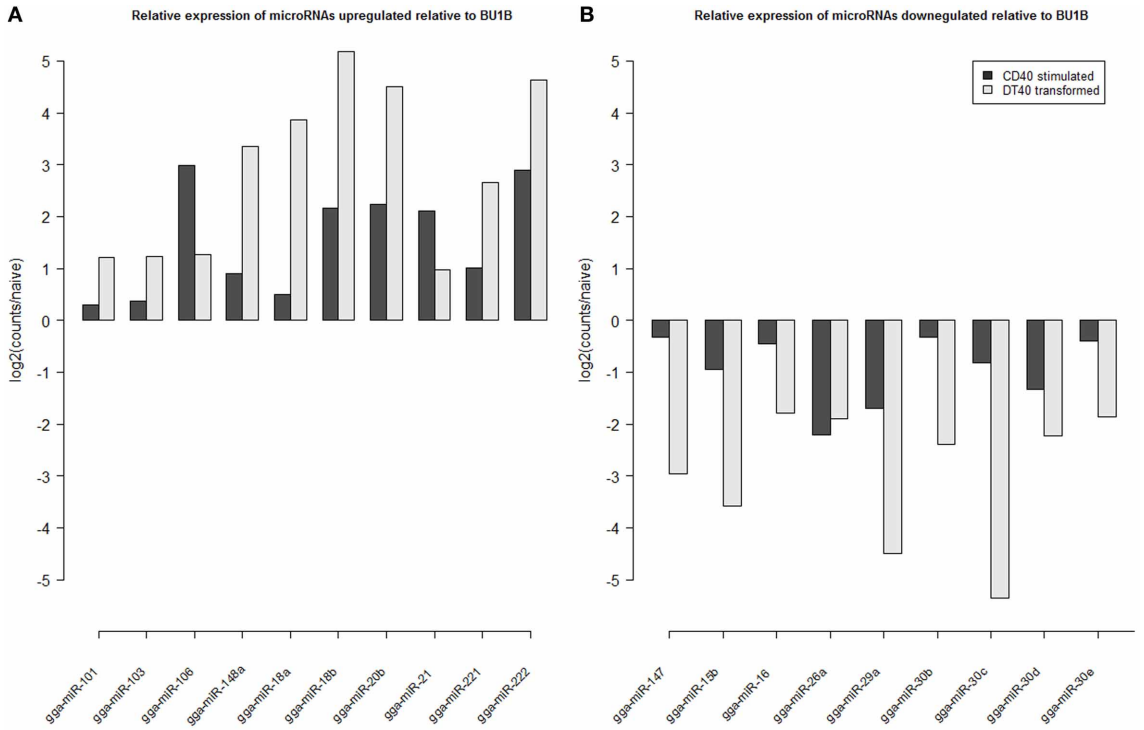


FIGURE 3 | Log2 fold change in expression of microRNAs (A) up-regulated in CD40L stimulated (StimB) and DT40 transformed cells (DT40) relative to naïve B cells (BU1B); and (B) down-regulation in CD40L stimulated (StimB) and DT40 transformed cells (DT40) relative to naïve B cells (BU1B).

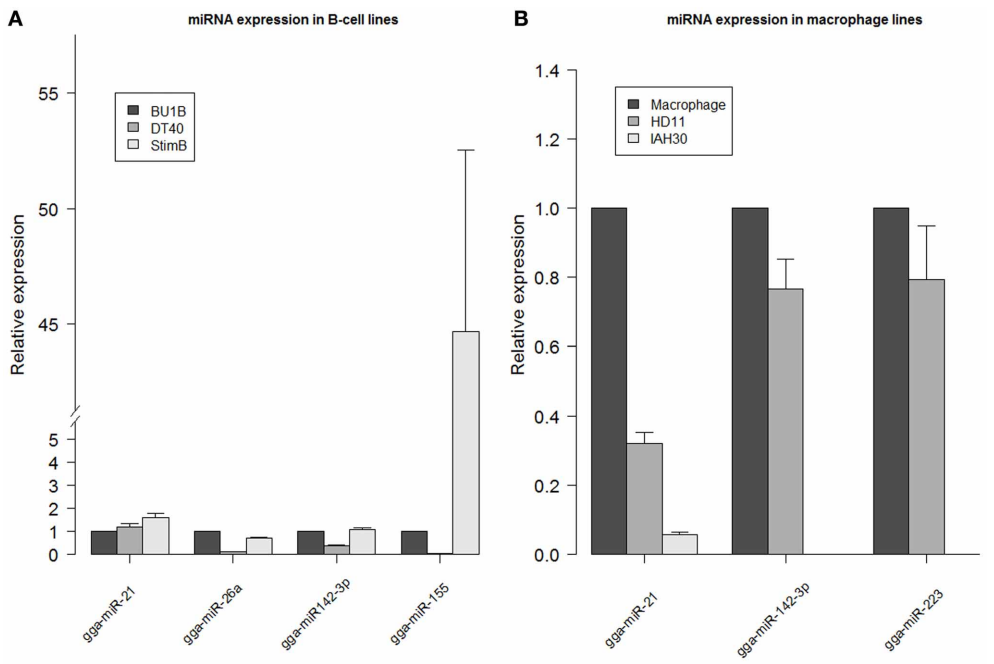


FIGURE 4 | miRNA expression levels determined by qRT-PCR. Relative expression of gga-miR-21, gga-miR-26a, gga-miR142-3p and gga-miR-155 measured in RNA extracted from DT40 and StimB cells compared to Bu1B cells **(A)** and gga-miR142-3p and gga-miR-223 in HD11 and IAH30 compared to chicken macrophages **(B)**. Results represent the mean of triplicate assays with error bars showing the standard error of the mean. Data normalized to gga-let-7a. Note the broken axis in **(A)**.

observed for gga-miR-21, which accounted for 28.8% of all miRNAs expressed in these cells. The high level of gga-miR-21 was demonstrated in normal macrophages also (**Figure 4B**). Other miRNAs which are expressed at high levels in HD11 include gga-miR-142-3p (10.5%), gga-miR-223 (6.9%), gga-miR-19b (4%), gga-miR-20a (3.7%) and gga-miR-22 (3.4%).

More than half (53%) of the total 175,408,236 sequences from the IAH30 turkey cell line mapped to known mature chicken miRNA sequences. Of those, a large majority matched with the gga-miR-21 (33.4%). Other chicken miRNAs that are expressed at high levels in IAH30 cells include gga-miR-24 (5.5%), gga-miR-27b (4%), gga-miR-19b (3.9%), gga-miR-20a (3.9%), gga-miR-148a (3.7%), gga-miR-23b (3%), and gga-miR-92 (3%). Further studies are required to identify the functional significance of these miRNAs and further characterize the other turkey miRNAs. Interestingly, gga-miR-142-3p and gga-miR-223 were significantly downregulated in IAH30 cells compared to HD11 cells **Figure 4** For gga-miR-142-3p, there are only 19 reads in IAH30 compared to 75,670 reads in HD11. Similarly, there are 10 and 50,275 reads representing gga-miR-223 in IAH30 and HD11 respectively. This difference is further confirmed by qRT-PCR (**Figure 4B**).

miRNA EXPRESSION IN cESC LINE BP25

Sequencing from cell line BP25 gave a total of 1,624,435 reads out of which 1,146,503 (70.6%) passed QC and mapped to known mature chicken miRNAs. The most predominant miRNA population expressed in BP25 was indeed the ES cell-specific miRNAs belonging to the miR-302-367 cluster. The expression levels of five miRNAs in the miR-302-367 cluster (miR-302a, miR-302b, miR-302c, miR-302d, and miR-367) accounted for 39.5% of all the sequenced miRNAs. Another miRNA expressed at high levels in the BP25 cell line is gga-miR-21 that accounted for 23.3% of the miRNAome. In addition, miR-17-92 cluster of miRNAs (miR-17, miR-18a, miR-19a, miR-20a, miR-19b-1, and miR-92-1) was also expressed at relatively high levels, accounting for 12.4% of the miRNAome in BP25.

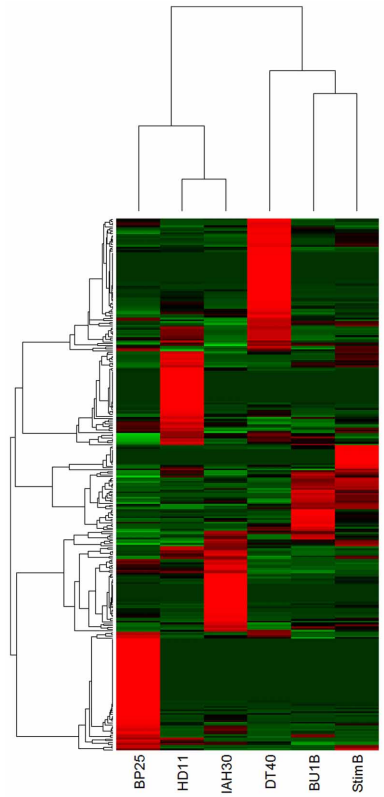


FIGURE 5 | Heatmap demonstrating differential clustering of the miRNAs from avian cell types based on expression levels calculated from normalized read counts (Red indicates high levels of expression and green to indicate low levels). BP25 is the cESC line BP25 (Pain et al., 1996); BU1B is the naïve embryonic B lymphocyte; DT40 is the data from DT40 cells (Buerstedde et al., 2002); stimB is the CD40L-induced B-cells; HD11 and IAH30 represent HD11 (Beug et al., 1979) and IAH30 (Lawson et al., 2001) cell lines respectively.

COMPARISON OF THE miRNA PROFILES OF DIFFERENT AVIAN CELL TYPES

Comparison of the differential expression of miRNA, based on the normalized counts, showed clustering of miRNAs in various avian cell types (**Figure 5**). For example, cESC line BP25 clearly demonstrated clustering of highly expressed ES cell-specific miRNAs, which are not expressed in any of the other cell lines. Similarly, DT40 cells showed a distinct profile of miRNA expression. The two cell lines IAH30 and HD11 coming from different species (turkey and chicken respectively), clearly showed distinct expression profiles despite being of macrophage origin. Naïve and stimulated B cells showed clustering based on miRNA expression profiles, yet demonstrated specific miRNA expression patterns, with a group of miRNAs showing high expression in stimulated B cells that are not present in the other cell lines.

DISCUSSION

Deep sequencing using Illumina platform can be valuable for obtaining miRNAome data, and we have used this for determining the miRNA expression levels from cell lines of chicken, a model avian species. Comparison of the normalized read counts was used to obtain digital data on expression levels of individual, already known mature miRNAs.

As in mammals, B-lymphocytes in birds are one of the critical components of the immune system responsible for the production of antibodies to specific antigens, pathogens and vaccines. Chickens have a distinct organ called BF in which the naïve B cells mature before being exported to the periphery. Naïve B cells are activated through a combination of signals from the antigen and through the binding of the CD40 ligand to CD40 which drives proliferation. The miRNAs which showed significant increase upon CD40L-stimulation included gga-miR-21, gga-miR-155, gga-miR-146a, gga-miR-20b, gga-miR-106, gga-miR-222 and gga-miR-22 (**Figure 2**). Some of these miRNAs have already been well-documented for their roles in cell proliferation and cancer. The highly expressed gga-miR-155 has been extensively studied and shown to be associated with cell proliferation in a number of cancers, as well as in autoimmunity (Leng et al., 2011; Wang and Wu, 2012). Interestingly, miR-155 was first discovered

in the chicken as part of the c-*bic* transcript in ALV-transformed lymphomas (Clurman and Hayward, 1989). High expression of miR-155 upon CD40L stimulation is consistent the major role of this miRNA in proliferation. Other viral oncogenes such as v-Rel have also been shown to drive miR-155 expression (Bolisetty et al., 2009). As v-Rel is an NF-κB homolog, it is possible that the increased expression of miR-155 by CD40L is mediated through the NF-κB pathway, although other signaling systems may also be involved. Another miRNA that shows increase in expression after CD40L-stimulated B-cells is gga-miR-146a. As a multifaceted miRNA, its role in hematopoiesis, immune response and cancer has been well documented (Labbaye and Testa, 2012). Activation of miR-146a, thought to be through the NF-κB pathway, has also shown to be important in the innate immune responses (Williams et al., 2008). Our study demonstrating the upregulation of miR-146a through CD40L interaction further adds to our understanding of the molecular pathways of biogenesis of miR-146a and CD40L functions.

A number of miRNAs were also down-regulated after CD40L stimulation. This included gga-miR-26a which showed a 4.5-fold decrease in expression. Interleukin-2 (IL-2) is essential for the growth and proliferation of T-cells (Cantrell and Smith, 1984), and we have previously shown the downregulation of miR-26a in MDV-transformed cell lines, where the decreased expression relieved its suppressive effect on the interleukin-2, potentially allowing proliferation (Xu et al., 2010). It is possible that the CD40L-stimulation also makes use of a similar pathway. The most down-regulated miRNA, which decreased in expression by almost 10-fold, was gga-miR-2954. This miRNA has only been reported in birds (*Gallus gallus* and *Taeniopygia guttata*) and was originally identified as being male-specific during early chick development (Zhao et al., 2010). Another avian-specific miRNA, gga-miR-1729 was also down-regulated, a miRNA originally discovered in developing chick embryo (Glazov et al., 2008). The change in expression after CD40L stimulation suggests that these miRNAs have roles beyond early chick development. Other down-regulated miRNAs include members of the miR-30 family of miRNAs (gga-miR-30c, gga-miR-30d, gga-miR-30a-5p) many of which have been implicated in a wide range of cancers (Gaziel-Sovran et al., 2011; Baraniskin et al., 2012; Cheng et al., 2012). Suppression of these miRNAs may therefore be related to cell division and proliferation of the B cells after stimulation with CD40L. By comparing the miRNA expression in naïve and CD40L-activated B-cells, we were able to identify changes in miRNA expression related to the CD40L-induced proliferation.

ALV-transformed cell line DT40 is extensively used in molecular genetic studies because it has high levels of recombination, making it a very useful system for *in vitro* gene knock out studies (Buerstedde et al., 2002). Although there have been extensive studies on important areas of cell biology using the DT40 cell system, there is only limited understanding of the miRNA expression and functions in this cell line. Global miRNA expression profiles of DT40 cells showed changes in the expression of a number of miRNAs (**Figure 2**). A number of miRNAs upregulated in DT40 cells also showed increased expression in CD40L-stimulated cells, suggesting that these miRNAs have a major role in cell proliferation. However, other miRNAs such as gga-miR-100 was up-regulated mainly in DT40 cells suggesting a more important role in transformation. Interestingly, miR-100 has been shown to be involved in a number of cancers in humans (Jung et al., 2011; Li et al., 2011; Oliveira et al., 2011; De Oliveira et al., 2012). On the other hand, gga-miR-146a, which was upregulated in CD40L stimulated B cells, is down-regulated in DT40 cells, providing further evidence for its role in the immune system.

A number of miRNAs down-regulated in DT40, including gga-miR-16, -30e, -30d, -30b, -30c, -26a, -147, -15b, and -29a, were also downregulated in CD40L-stimulated cells, suggesting conserved functions of cell division and proliferation. However, perhaps the most striking change is in gga-miR-155, which is the most up-regulated miRNA in CD40L stimulated B cells, but it is the most down-regulated miRNA in ALV transformed DT40 cells. One of the well characterized targets of miR-155, the transcription factor PU.1 is expressed in DT40 cells and has been shown to be important in the activation-induced cytidine deaminase (AID) expression and function (Luo and Tian, 2010). AID is important for B-cells to produce and maintain antibody diversity (Muramatsu et al., 2000). Reduced levels of miR-155 may help in maintaining high levels of PU.1, as these two have been shown to demonstrate inverse correlation in their expression (Thompson et al., 2011). Although it would be inaccurate to make direct comparisons of the miRNA profiles of DT40 cell lines and the naïve B cells stimulated with CD40L because of the significant differences between the two populations, the findings from the present study suggest the changes in the expression of some of these miRNAs could be contributing to the proliferative and unique recombinogenic properties of DT40 cells.

Macrophages play primary roles in both innate and adaptive immune responses, and *in vitro* studies on their function using macrophage cell lines have provided significant understanding on such responses. Avian myelocytomatosis virus (MC29)-transformed chicken macrophage cell line HD11 (Beug et al., 1979) is widely used in examining the innate immune functions. Although a number of miRNAs have been implicated in modulating innate immune functions, the miRNA profiles of these cells have not been examined. Although significant further studies are required to obtain the digital miRNA expression data of these cells, our data provide a snapshot of the expression profiles of the miRNAs that may have relevance to studies on their function. Among all the miRNAs expressed in these cells, gga-miR-21 alone accounted for nearly a third (28.8%). However, miR-21 is highly expressed in normal macrophages as well as the cESC line BP25. The data presented here could be valuable in examining the changes in miRNA expression profiles following the *in vitro* activation of specific signaling pathways, for which these cell types are widely used.

IAH30 is a turkey macrophage cell line (Lawson et al., 2001) transformed by the acutely transforming ALV subgroup J 966 virus with a transduced v-*myc* oncogene (Chesters et al., 2001). Although there has been progress in sequencing of the turkey genome (Dalloul et al., 2010), the annotation of the miRNAs is still not complete and no mature turkey miRNA sequences are available in miRBase. Hence we have used the deep sequencing data from IAH30 to examine the changes in the expression of turkey homologs of chicken miRNAs.

The two miRNAs that are significantly downregulated in IAH30 cells compared to HD11 –gga-miR-142-3p and

gga-miR-223 (**Figure 4B**) have been shown to be involved in haemopoietic cell proliferation (Sun et al., 2010) and macrophage differentiation (Ismail et al., 2013). The vastly different abundances of these two key miRNAs suggest potential differences between the IAH30 and HD11 macrophage cell lines which may have implications for experiments which utilize them.

Analyzis of the miRNA profiles in the BP25 cESC line showed that the ES cell-specific miRNAs belonging to the miR-302-367 cluster were the most predominant miRNA population expressed in BP25. The expression levels of five miRNAs in the miR-302-367 cluster (miR-302a, miR-302b, miR-302c, miR-302d, and miR-367) accounted for 39.5% of all the sequenced miRNAs. This cluster of miRNAs is also expressed at high levels in human ESCs (Gunaratne, 2009), suggesting conserved functions of these miRNAs in mammals and birds. It has also been demonstrated in mammals that only a small subset of miRNAs, mostly seen as clusters in the genome, are expressed in the ES cells. For example, over 75% of the miRNAs expressed in mouse ES cells are represented by 6 loci (Calabrese et al., 2007). More recently, the miR-302-367 cluster was found to be highly expressed in differentiated blastoderm and primordial cells (Lee et al., 2011). Although shown to be significantly induced in human embryonal carcinoma cells (Suh et al., 2004), none of the miRNAs from the miR-302-367 cluster were detected in any of the other cell lines examined in this study, suggesting the suppression of these miRNAs upon differentiation. miRNAs belonging to the miR-302-367 cluster have been shown to regulate cell growth, metabolism, transcription (Dyce et al., 2010) and chromatin modification (Ren et al., 2009), demonstrating the potential importance of these miRNAs in maintaining the stem cell phenotype. BP25 cell line also showed high levels of expression of gga-miR-21 accounting for 23.3% of the miRNAome. Unlike in the BP25 cESC line, miR-21 levels are low in mammalian ES cells (Gunaratne, 2009), although the potential functional significance is not known. However, increased miR-21 expression is not unique to the chicken ES cell line, as high levels of expression of miR-21 have been detected in a number of human cancer cell lines (Slaby et al., 2007; Jiang et al., 2008). The miR-17-92 cluster of miRNAs (miR-17, miR-18a, miR-19a, miR-20a, miR-19b-1, and miR-92-1) was also expressed at relatively high levels (accounting for 12.4% of the miRNAome) in the BP25 cell line. As multifaceted miRNAs, these are not specific to the ESC, and have been associated with a number of cancers (Krichevsky and Gabriely, 2009), therefore the role of the miR-17-92 cluster in the BP25 cESC line may be related to its proliferative functions.

Taken together, we present the expression profiles of miRNAs determined by deep sequence analysis of the small RNA population from a number of avian cell types under conditions such as CD40L stimulation of B cells. The study also provides a snapshot of the miRNA profiles of cell lines such as DT40, HD11 and IAH30, transformed by the activation/transduction of *myc* oncogene. Although additional studies are required for precise characterization of the changes in miRNA expression and the functional significance of these changes, this study provides insights into the potential roles of miRNAs in the hematopoietic lineages of cells in a model non-mammalian species.

ACKNOWLEDGMENTS

This project was supported by the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/J004243/1; BB/J004235/1; BB/I01361X/1; BB/I014284/1). We thank Dr. Bertrand Pain, Pluripotency and Differentiation control of Embryonic Stem cells group, INSERM, Lyon for providing the BP25 cell line.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Non-Coding_RNA/10.3389/fgene.2013.00153/abstract

REFERENCES

Bachl, J., Caldwell, R. B., and Buerstedde, J. M. (2007). Biotechnology and the chicken B cell line DT40. *Cytogenet. Genome Res.* 117, 189–194. doi: 10.1159/000103179

Baraniskin, A., Birkenkamp-Demtroder, K., Maghnouj, A., Zollner, H., Munding, J., Klein-Scory, S., et al. (2012). MiR-30a-5p suppresses tumor growth in colon carcinoma by targeting DTL. *Carcinogenesis* 33, 732–739. doi: 10.1093/carcin/bgs020

Beug, H., Von Kirchbach, A., Doderlein, G., Conscience, J. F., and Graf, T. (1979). Chicken hematopoietic cells transformed by seven strains of defective avian leukemia viruses display three distinct phenotypes of differentiation. *Cell* 18, 375–390. doi: 10.1016/0092-8674(79)90057-6

Bolisetty, M. T., Dy, G., Tam, W., and Beemon, K. L. (2009). Reticuloendotheliosis virus strain T induces miR-155, which targets JARID2 and promotes cell survival. *J. Virol.* 83, 12009–12017. doi: 10.1128/JVI.01182-09

Buerstedde, J. M., Arakawa, H., Watahiki, A., Carninci, P. P., Hayashizaki, Y. Y., Korn, B., et al. (2002). The DT40 web site: sampling and connecting the genes of a B cell line. *Nucleic Acids Res.* 30, 230–231. doi: 10.1093/nar/30.1.230

Burnside, J., Ouyang, M., Anderson, A., Bernberg, E., Lu, C., Meyers, B. C., et al. (2008). Deep sequencing of chicken microRNAs. *BMC Genomics* 9:185. doi: 10.1186/1471-2164-9-185

Calabrese, J. M., Seila, A. C., Yeo, G. W., and Sharp, P. A. (2007). RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* 104, 18097–18102. doi: 10.1073/pnas.0709193104

Cantrell, D. A., and Smith, K. A. (1984). The interleukin-2 T-cell system: a new cell growth model. *Science* 224, 1312–1316. doi: 10.1126/science.6427923

Cheng, C. W., Wang, H. W., Chang, C. W., Chu, H. W., Chen, C. Y., Yu, J. C., et al. (2012). MicroRNA-30a inhibits cell migration and invasion by downregulating vimentin expression and is a potential prognostic marker in breast cancer. *Breast Cancer Res. Treat.* 134, 1081–1093. doi: 10.1007/s10549-012-2034-4

Chesters, P. M., Howes, K., Mckay, J. C., Payne, L. N., and Venugopal, K. (2001). Acutely transforming avian leukosis virus subgroup J strain 966: defective genome encodes a 72-kilodalton Gag-Myc fusion protein. *J. Virol.* 75, 4219–4225. doi: 10.1128/JVI.75.9.4219-4225.2001

Clurman, B. E., and Hayward, W. S. (1989). Multiple proto-oncogene activations in avian leukosis virus-induced lymphomas: evidence for stage-specific events. *Mol. Cell. Biol.* 9, 2657–2664.

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Blomberg Le, A., et al. (2010). Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol* 8:e1000475. doi: 10.1371/journal.pbio.1000475

De Oliveira, J. C., Scrideli, C. A., Brassesco, M. S., Morales, A. G., Pezuk, J. A., Queiroz Rde, P., et al. (2012). Differential miRNA expression in childhood acute lymphoblastic leukemia and association with clinical and biological features. *Leuk. Res.* 36, 293–298. doi: 10.1016/j.leukres.2011.10.005

Dyce, P. W., Toms, D., and Li, J. (2010). Stem cells and germ cells: microRNA and gene expression signatures. *Histol. Histopathol.* 25, 505–513.

Friedlander, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., et al. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* 26, 407–415. doi: 10.1038/nbt1394

Gaziel-Sovran, A., Segura, M. F., Di Micco, R., Collins, M. K., Hanniford, D., Vega-Saenz De Miera, E., et al. (2011). miR-30b/30d regulation of GalNAc transferases enhances invasion and immunosuppression during metastasis. *Cancer Cell* 20, 104–118. doi: 10.1016/j.ccr.2011.05.027

Glazov, E. A., Cottee, P. A., Barris, W. C., Moore, R. J., Dalrymple, B. P., and Tizard, M. L. (2008). A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. *Genome Res.* 18, 957–964. doi: 10.1101/gr.074740.107

Gunaratne, P. H. (2009). Embryonic stem cell microRNAs: defining factors in induced pluripotent (iPS) and cancer (CSC) stem cells. *Curr. Stem. Cell Res. Ther.* 4, 168–177. doi: 10.2174/157488809789057400

Ismail, N., Wang, Y., Dakhlallah, D., Moldovan, L., Agarwal, K., Batte, K., et al. (2013). Macrophage microvesicles induce macrophage differentiation and miR-223 transfer. *Blood* 121, 984–995. doi: 10.1182/blood-2011-08-374793

Jiang, J., Gusev, Y., Aderca, I., Mettler, T. A., Nagorney, D. M., Brackett, D. J., et al. (2008). Association of MicroRNA expression in hepatocellular carcinomas with hepatitis infection, cirrhosis, and patient survival. *Clin. Cancer Res.* 14, 419–427. doi: 10.1158/1078-0432.CCR-07-0523

Jung, D. E., Wen, J., Oh, T., and Song, S. Y. (2011). Differentially expressed microRNAs in pancreatic cancer stem cells. *Pancreas* 40, 1180–1187. doi: 10.1097/MPA.0b013e318221b33e

Kothlow, S., Morgenroth, I., Tregaskes, C. A., Kaspers, B., and Young, J. R. (2008). CD40 ligand supports the long-term maintenance and differentiation of chicken B cells in culture. *Dev. Comp. Immunol.* 32, 1015–1026. doi: 10.1016/j.dci.2008.01.012

Krichevsky, A. M., and Gabriely, G. (2009). miR-21: a small multi-faceted RNA. *J. Cell Mol. Med.* 13, 39–53. doi: 10.1111/j.1582-4934.2008.00556.x

Labbaye, C., and Testa, U. (2012). The emerging role of MIR-146A in the control of hematopoiesis, immune function and cancer. *J. Hematol. Oncol.* 5, 13. doi: 10.1186/1756-8722-5-13

Lawson, S., Rothwell, L., Lambrecht, B., Howes, K., Venugopal, K., and Kaiser, P. (2001). Turkey and chicken interferon-gamma, which share high sequence identity, are biologically cross-reactive. *Dev. Comp. Immunol.* 25, 69–82. doi: 10.1016/S0145-305X(00)00044-6

Lee, S. I., Lee, B. R., Hwang, Y. S., Lee, H. C., Rengaraj, D., Song, G., et al. (2011). MicroRNA-mediated posttranscriptional regulation is required for maintaining undifferentiated properties of blastoderm and primordial germ cells in chickens. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10426–10431. doi: 10.1073/pnas.1106141108

Leng, R. X., Pan, H. F., Qin, W. Z., Chen, G. M., and Ye, D. Q. (2011). Role of microRNA-155 in autoimmunity. *Cytokine Growth Factor Rev.* 22, 141–147. doi: 10.1016/j.cytogfr.2011.05.002

Li, B. H., Zhou, J. S., Ye, F., Cheng, X. D., Zhou, C. Y., Lu, W. G., et al. (2011). Reduced miR-100 expression in cervical cancer and precursors and its carcinogenic effect through targeting PLK1 protein. *Eur. J. Cancer* 47, 2166–2174. doi: 10.1016/j.ejca.2011.04.037

Luo, H., and Tian, M. (2010). Transcription factors PU.1 and IRF4 regulate activation induced cytidine deaminase in chicken B cells. *Mol. Immunol.* 47, 1383–1395. doi: 10.1016/j.molimm.2010.02.016

Merkerova, M., Belickova, M., and Bruchova, H. (2008). Differential expression of microRNAs in hematopoietic cell lineages. *Eur. J. Haematol.* 81, 304–310. doi: 10.1111/j.1600-0609.2008.01111.x

Morgan, R. W., and Burnside, J. (2011). Roles of avian herpesvirus microRNAs in infection, latency, and oncogenesis. *Biochim. Biophys Acta* 1809, 654–659. doi: 10.1016/j.bbagrm.2011.06.001

Mortazavi, A., W. B., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 102, 553–563. doi: 10.1016/S0092-8674(00)00078-7

Oliveira, J. C., Brassesco, M. S., Morales, A. G., Pezuk, J. A., Fedatto, P. F., Da Silva, G. N., et al. (2011). MicroRNA-100 acts as a tumor suppressor in human bladder carcinoma 5637 cells. *Asian Pac. J. Cancer Prev.* 12, 3001–3004.

Pain, B., Clark, M. E., Shen, M., Nakazawa, H., Sakurai, M., Samarut, J., et al. (1996). Long-term *in vitro* culture and characterisation of avian embryonic stem cells with multiple morphogenetic potentialities. *Development* 122, 2339–2348.

Ramkissoon, S. H., Mainwaring, L. A., Ogasawara, Y., Keyvanfar, K., Mccoy, J. P. Jr., Sloand, E. M., et al. (2006). Hematopoietic-specific microRNA expression in human cells. *Leuk. Res.* 30, 643–647. doi: 10.1016/j.leukres.2005.09.001

Ren, J., Jin, P., Wang, E., Marincola, F. M., and Stroncek, D. F. (2009). MicroRNA and gene expression patterns in the differentiation of human embryonic stem cells. *J. Transl. Med.* 7, 20. doi: 10.1186/1479-5876-7-20

Rothwell, C. J., Vervelde, L., and Davison, T. F. (1996). Identification of chicken Bu-1 alloantigens using the monoclonal antibody AV20. *Vet. Immunol. Immunopathol.* 55, 225–234. doi: 10.1016/S0165-2427(96)05635-8

Slaby, O., Svoboda, M., Fabian, P., Smerdova, T., Knoflickova, D., Bednarikova, M., et al. (2007). Altered expression of miR-21, miR-31, miR-143 and miR-145 is related to clinicopathologic features of colorectal cancer. *Oncology* 72, 397–402. doi: 10.1159/000113489

Suh, M. R., Lee, Y., Kim, J. Y., Kim, S. K., Moon, S. H., Lee, J. Y., et al. (2004). Human embryonic stem cells express a unique set of microRNAs. *Dev. Biol.* 270, 488–498. doi: 10.1016/j.ydbio.2004.02.019

Sun, W., Shen, W., Yang, S., Hu, F., Li, H., and Zhu, T. H. (2010). miR-223 and miR-142 attenuate hematopoietic cell proliferation, and miR-223 positively regulates miR-142 through LMO2 isoforms and CEBP-beta. *Cell Res.* 20, 1158–1169. doi: 10.1038/cr.2010.134

Thompson, R. C., Herscovitch, M., Zhao, I., Ford, T. J., and Gilmore, T. D. (2011). NF-kappaB down-regulates expression of the B-lymphoma marker CD10 through a miR-155/PU.1 pathway. *J. Biol. Chem.* 286, 1675–1682. doi: 10.1074/jbc.M110.177063

Tregaskes, C. A., Glansbeek, H. L., Gill, A. C., Hunt, L. G., Burnside, J., and Young, J. R. (2005). Conservation of biological properties of the CD40 ligand, CD154 in a non-mammalian vertebrate. *Dev. Comp. Immunol.* 29, 361–374. doi: 10.1016/j.dci.2004.09.001

Wang, J., and Wu, J. (2012). Role of miR-155 in breast cancer. *Front. Biosci.* 17, 2350–2355. doi: 10.2741/4056

Williams, A. E., Perry, M. M., Moschos, S. A., Larner-Svensson, H. M., and Lindsay, M. A. (2008). Role of miRNA-146a in the regulation of the innate immune response and cancer. *Biochem. Soc. Trans.* 36, 1211–1215. doi: 10.1042/BST0361211

Xu, H., Yao, Y., Smith, L. P., and Nair, V. (2010). MicroRNA-26a-mediated regulation of interleukin-2 expression in transformed avian lymphocyte lines. *Cancer Cell Int.* 10, 15. doi: 10.1186/1475-2867-10-15

Yao, Y., Smith, L. P., Petherbridge, L., Watson, M., and Nair, V. (2012). Novel microRNAs encoded by duck enteritis virus. *J. Gen. Virol.* 93, 1530–1536. doi: 10.1099/vir.0.040634-0

Yao, Y., Zhao, Y., Smith, L. P., Lawrie, C. H., Saunders, N. J., Watson, M., et al. (2009). Differential expression of microRNAs in Marek's disease virus-transformed T-lymphoma cell lines. *J. Gen. Virol.* 90, 1551–1559. doi: 10.1099/vir.0.009902-0

Yao, Y., Zhao, Y., Xu, H., Smith, L. P., Lawrie, C.H., Watson, M., et al. (2008). MicroRNA profile of Marek's disease virus-transformed T-cell line MSB-1: predominance of virus-encoded microRNAs. *J. Virol.* 82, 4007–4015. doi: 10.1128/JVI.02659-07

Zhao, D., Mcbride, D., Nandi, S., Mcqueen, H. A., Mcgrew, M. J., Hocking, P. M., et al. (2010). Somatic sex identity is cell autonomous in the chicken. *Nature* 464, 237–242. doi: 10.1038/nature08852

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 April 2013; accepted: 22 July 2013; published online: 14 August 2013.
Citation: Yao Y, Charlesworth J, Nair V and Watson M (2013) MicroRNA expression profiles in avian haemopoietic cells. Front. Genet. 4:153. doi: 10.3389/fgene.2013.00153
This article was submitted to Frontiers in Non-Coding RNA, a specialty of Frontiers in Genetics.

JVI Journals.ASM.org

An Avian Retrovirus Uses Canonical Expression and Processing Mechanisms To Generate Viral MicroRNA

Yongxiu Yao,[a] Lorraine P. Smith,[a] Venugopal Nair,[a] Mick Watson[b]

Viral Oncogenesis Group, The Pirbright Institute, Compton Laboratory, Compton, Berkshire, United Kingdom[a]; Ark-Genomics, The Roslin Institute, R(D)SVS, University of Edinburgh, Division of Genetics and Genomics, Easter Bush, Midlothian, United Kingdom[b]

To date, the vast majority of known virus-encoded microRNAs (miRNAs) are derived from polymerase II transcripts encoded by DNA viruses. A recent demonstration that the bovine leukemia virus, a retrovirus, uses RNA polymerase III to directly transcribe the pre-miRNA hairpins to generate viral miRNAs further supports the common notion that the canonical pathway of miRNA biogenesis does not exist commonly among RNA viruses. Here, we show that an exogenous virus-specific region, termed the E element or XSR, of avian leukosis virus subgroup J (ALV-J), a member of avian retrovirus, encodes a novel miRNA, designated E (XSR) miRNA, using the canonical miRNA biogenesis pathway. Detection of novel microRNA species derived from the E (XSR) element, a 148-nucleotide noncoding RNA with hairpin structure, showed that the E (XSR) element has the potential to function as a microRNA primary transcript, demonstrating a hitherto unknown function with possible roles in myeloid leukosis associated with ALV-J.

Retroviruses are a large group of enveloped viruses associated with a variety of diseases in a wide range of host species. Avian retroviruses, the Rous sarcoma virus (RSV) and avian leukosis virus (ALV), are historically known for their ability to induce a number of types of cancer in poultry (1). In addition to their pathogenic roles, retroviruses have provided significant insights into transcriptional regulation in a cell-type-specific manner (2). The retroviral genome includes a number of *cis*-acting elements (3). One such element is the 148-nucleotide exogenous virus-specific region (E or XSR) identified in the SR-A and Pr-C strains of RSV at the 5′ and the 3′ sides of the *src* gene, respectively (4, 5). The functions of the E (XSR) element are not clear although requirement of a 400-nucleotide region that included the E (XSR) element for oncogenicity of the recombinant avian retrovirus NTRE7 has been shown (6). The E (XSR) sequence exhibits several unusual features; it has a noncoding RNA capable of forming characteristic hairpin structures (7). From its location at two different sites on either side of the *src* gene in the two RSV strains, it is clear that the functions of E (XSR) can be exerted over distance. Based on these observations, it was speculated that E (XSR) may function as a transcriptional enhancer (5).

Interest in the E (XSR) element was revived when it was demonstrated in the 3′ noncoding region of the genome of HPRS-103, the ALV subgroup J (ALV-J) prototype virus (8), identified in the United Kingdom in 1988 as the causative agent of myeloid leukosis, which rapidly became a worldwide health and welfare problem in chickens (9–13). The E (XSR) sequence is conserved in the majority of the ALV-J isolates although deletions or modifications in this sequence have also been seen (13–16). The role of the E (XSR) element in the pathobiology of ALV-J is not known although potential C/EBP and c-Ets-1 binding sites have been predicted in the sequence (13, 14). However, ALV-J strains with deletions or mutations in the E (XSR) element have also been isolated from clinical cases (9–11, 14, 17). Our previous studies using HPRS-103 clones with precise deletions in the E (XSR) element indicated that these elements are essential for oncogenicity, but this was related to the genetic background of the birds (7). Despite the presence of the E (XSR) element and its association with the oncogenicity of RSV and ALV-J, the molecular mechanisms of the E (XSR) element functions remain unclear. Although an enhancer-like function has been speculated (5, 14), firm supporting evidence is still lacking.

In many organisms, including several viruses, microRNAs (miRNAs) are well recognized as major regulators of gene expression (18). Given their profound ability to regulate multiple targets, these molecules are exploited particularly by several DNA viruses as tools for manipulating the cellular environment (19, 20). RNA viruses are generally thought not to contain pre-miRNA structures to avoid endonuclease-mediated cleavage of the genome, antigenome, and mRNAs. Although retroviruses have not been widely documented to exploit the miRNA pathway (21), a recent demonstration of a conserved cluster of RNA polymerase III (Pol III)-transcribed miRNAs from the bovine leukemia virus (BLV) genome (22, 23) showed the potential of retroviruses to encode miRNAs. The E (XSR) element sequences from ALV-J strains show hairpin-like structures suggestive of miRNA precursors although the existence of any mature miRNA has not been demonstrated in ALV-J-infected/transformed cells. Using a deep-sequencing approach on one of the ALV-J-transformed cell lines, we identified a novel small-RNA population encoded from within the E (XSR) element.

MATERIALS AND METHODS

Cells. HEK293T cells and the chicken embryo fibroblast (CEF) cell line DF-1 (24) were maintained in Dulbecco's modified Eagle's medium

Received 7 October 2013 Accepted 7 October 2013

Published ahead of print 23 October 2013

Address correspondence to Venugopal Nair, venugopal.nair@pirbright.ac.uk.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.02921-13.

doi:10.1128/JVI.02921-13

(DMEM) supplemented with 10% fetal calf serum (FCS) (Sigma). A reticuloendotheliosis virus T (REV-T)-transformed turkey spleen cell (TSC) line, AVOL-1T, and IAH30, a turkey macrophage (MΦ) cell line (25) transformed by the acutely transforming ALV subgroup J 966 virus with a transduced v-*myc* oncogene (26), were grown at 38.5°C in 5% CO_2 in RPMI 1640 medium containing 10% FCS, 2% chicken serum, 10% tryptose phosphate broth, 0.1% 2-mercaptoethanol, and 1% sodium pyruvate. AVO4-1B3 cells, an avian blastoderm cell line transformed by acutely transforming ALV-J isolate 1B (27), was maintained in Eagle's minimal essential medium (EMEM) supplemented with 10% FCS. Turkey spleen cells were prepared from spleen tissues of uninfected turkey by using Histopaque-1083 (Sigma-Aldrich) density gradient centrifugation. Chicken macrophages were prepared from bone marrow of adult uninfected chickens using procedures described previously (12). Briefly, femoral bone marrow was flushed out with RPMI medium, and the contents were passed through a cell sieve. After Histopaque-1083 density gradient centrifugation, the cells from the interface were collected, washed twice using RPMI medium supplemented with 5% FCS and 5% chicken serum, plated into 60-mm dishes, and cultured at 38.5°C in 5% CO_2. The medium was changed after 3 days, and the cells were used for virus infection.

Plasmids. The miR-155 expression plasmid pEF6-Bic was described previously (28). The MHV-miR-M1-7-3p expression plasmid pIDTSmartKan-MHV68-M1-7 was kindly provided by Chris Sullivan, University of Texas, Austin, TX. The E (XSR) miRNA expression cassette was PCR amplified from HPRS-103 genomic DNA with primers 5′-TCG ATAGGAAGCTTAAAGCAGTGCATGGGTAGGGGT-3′ and 5′-TCAC GTAATCTAGACCACCTTACTTCCACCAATCGACG-3′. The isolated fragments were either cloned into pGEM-T-easy by TA cloning or into pcDNA3.1-His/myc via restriction enzyme sites HindIII and XbaI. A luciferase sensor reporter was constructed by synthesizing four tandem repeats of artificial target sites with a perfect match to E (XSR) miRNA as sense and antisense oligomers; the construct was annealed and cloned into the 3′ untranslated region (UTR) of the *Renilla* luciferase gene through the NotI-XhoI site in the psiCHECK-2 vector (Promega). A control mutant reporter construct (sensor-mu) was generated alongside the wild-type reporter construct to be the negative control. In the mutant construct, nucleotides 2, 4, 6, and 8 of the microRNA response element (MRE) were changed to different bases, thus preventing recognition of the target site by the miRNA. In all cases, the cloned sequences were confirmed by sequence analysis.

Deep sequencing of an IAH30 small-RNA library. RNA extraction for miRNA profiling was carried out as previously described (29) using an miRVana miRNA isolation kit (Ambion, United Kingdom). Sequencing of the miRNAs was carried out on an Illumina GAIIx platform by GATC Biotech. After sequencing, adaptor and primer/dimer sequences were removed, yielding approximately 1.5 million 36-bp single-end reads. Using the Novoalign short-read mapper (Novocraft, Selangor, Malaysia), we mapped the reads to the known chicken mature miRNAs downloaded from miRBase (www.miRBase.org) and the HPRS-103 genome sequence (GenBank accession number Z46390.1). To correct for differences in library size and sequencing depth, raw mapped-read counts were scaled to reads per kilobase per million sequenced reads (30).

Northern blotting. Total RNA was extracted from cultured cells with an miRNeasy Kit (Qiagen) according to the manufacturer's instructions. Samples of 20 μg of total RNA were resolved using a 15% polyacrylamide–1× Tris-borate–EDTA–8 M urea gel and blotted to a GeneScreen Plus membrane (Perkin-Elmer). The probe sequences of DNA oligonucleotides with sequences complementary to candidate miRNAs are as follows: miR-155 probe, 5′-CCCCTATCACGATTAGCATTAA-3′; E (XSR) miRNA probe, 5′-CAGAGGCAACTTGAATAGTCTA-3′; and MHV68-M1-7 probe, 5′-AATAAAGGTGGGCGCGATATC-3′. The 5.8S probe sequence is 5′-TTCTTCATCGACGCACGAGC-3′. The probes were end labeled with [γ-^{32}P]ATP (Amersham) and T4 polynucleotide kinase (New England BioLabs) to generate high-specific-activity probes. Hybridization, washing, and autoradiography were carried out as previously described (31).

Dual-luciferase assay. The transfection of DF-1 cells and Vero cells was carried out with Lipofectamine 2000 (Invitrogen). CEFs were transfected with Lipofectamine (Invitrogen), and IAH30 cells were transfected with a Nucleofactor Transfection Kit T (Lonza). E (XSR) miRNA mimics were synthesized by Qiagen. Approximately 3×10^4 cells were seeded in each well of a 96-well plate. Except for the IAH30 cell line, which was transfected with reporter construct alone, DF-1, CEF, and Vero cells were cotransfected with 20 ng of each reporter construct in psiCHECK-2 vector along with either 100 ng of (E) XSR miRNA expression plasmid (DF-1) or a final concentration of 20 nM E (XSR) miRNA mimics (DF-1, CEF, and Vero cells) using different transfection reagents, as stated above, as per the manufacturer's protocols. In all cases, constitutively expressed firefly luciferase activity in the psiCHECK-2 vector served as a normalization control for transfection efficiency.

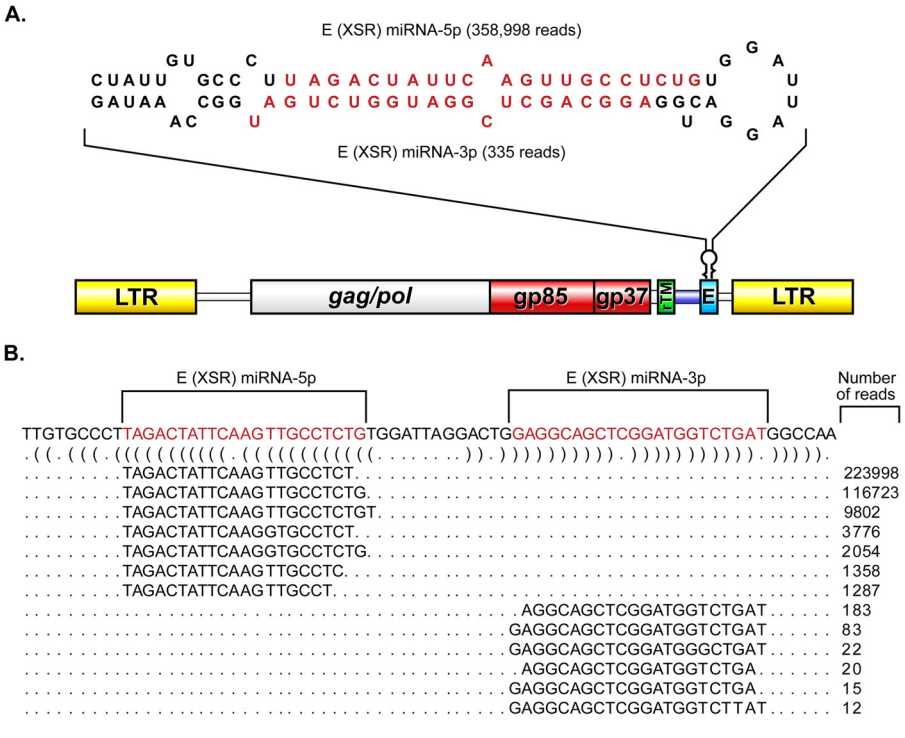
RNA polymerase II dependence assay. HEK293T cells in six-well plates were transfected with 2.5 μg of miRNA expression vector using Lipofectamine 2000 and, where indicated, were then treated 2 h later with a final concentration of 50 μg/ml α-amanitin (32). Total RNA was extracted at 24 h posttransfection, and Northern blot analysis was performed.

RNA interference (RNAi) assays. Silencer Select validated siRNAs against human Drosha (siRNA identification numbers [ID], s26491 and s26492, referred to as D1 and D2) and Dicer (siRNA ID s23754) were purchased from Ambion. They have been verified experimentally by the company in cell-based assays to reduce the expression of their individual target genes by 80% in at least three biological replicates. HEK293T cells in six-well plates were transfected with 20 nM Drosha or Dicer siRNA using Lipofectamine RNAi-MAX (Invitrogen) following the manufacturer's recommendations. At 24 h posttransfection, cells were cotransfected with 20 nM each siRNA and 2 μg of miRNA expression plasmid using Lipofectamine 2000. Twenty-four hours later, RNA was extracted, and Northern blot analysis was performed.

Stem-loop quantitative reverse transcription-PCR (qRT-PCR) for E (XSR) miRNA. Total RNA was extracted from cultured cells with an miRNeasy Kit (Qiagen) according to the manufacturer's instructions. miRNAs were quantified using custom TaqMan stem-loop microRNA assays (ABI) according to the manufacturer's recommendations using 10 ng of total RNA as a template for reverse transcription with the primer 5′-GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACG ACCAGAGG-3′, followed by quantitative PCR carried out using the forward primer 5′-CGGTCGACCTAGACTATTCAAGTTG-3′, reverse primer 5′-CAGTGCAGGGTCCGAGGT-3′, and probe 5′-TGGATACGA CCAGAGGC-3′. A TaqMan microRNA assay for let-7a (assay ID 000377) was purchased from ABI. Each reverse transcription reaction was performed twice independently, and each reaction mixture was used for triplicate PCR. The level of miRNA expression is presented as fold expression relative to the background amplification obtained with RNA isolated from either chicken MΦ (see Fig. 2B) or untransfected DF-1 (see Fig. 4A) and after normalization to the ubiquitously expressed cellular miRNA let-7a.

RESULTS

E (XSR) element encodes a novel miRNA. During miRNA profiling of an ALV-J-transformed turkey macrophage cell line IAH30 (25) by deep-sequence analysis on an Illumina GAIIx platform, we observed the presence of two distinct small RNA sequences that mapped perfectly to the E (XSR) element in the 3′ UTR of the HPRS-103 genome (GenBank accession number Z46390.1). The total reads of the two small RNA sequences account for 24.5% of the total IAH30 "miRNAome." Notably, only 580 reads were scattered across the 7,841-bp genome outside this region. The relative abundance (358,998 and 335 reads, respectively) (Fig. 1A and B) and size of these small RNAs suggested they are miRNA candi-



Name	Sequence	Reads
gga-miR-21	UAGCUUAUCAGACUGAUGUUGA	342,810
gga-miR-24	UGGCUCAGUUCAGCAGGAACAG	62,256
gga-miR-140	AGUGGUUUUACCCUAUGGUAG	52,589
gga-miR-140*	CCACAGGGUAGAACCACGGAC	51,777
gga-miR-19b	UGUGCAAAUCCAUGCAAAACUGA	39,417
gga-miR-20a	UAAAGUGCUUAUAGUGCAGGUAG	38,758
gga-miR-27b	UUCACAGUGGCUAAGUUCUGC	38,174
gga-miR-148a	UCAGUGCACUACAGAACUUUGU	36,199
gga-miR-23b	AUCACAUUGCCAGGGAUUACC	31,214
gga-miR-92	UAUUGCACUUGUCCCGGCCUG	29,912
gga-miR-19a	UGUGCAAAUCUAUGCAAAACUGA	28,501
gga-miR-29a	UAGCACCAUUUGAAAUCGGUU	21,331
gga-miR-20b	CAAAGUGCUCAUAGUGCAGGUAG	19,234
gga-miR-181a	AACAUUCAACGCUGUCGGUGAGU	18,770
gga-miR-26a	UUCAAGUAAUCCAGGAUAGGC	18,324
gga-miR-101	GUACAGUACUGUGAUAACUGAA	17,388
gga-miR-30a-5p	UGUAAACAUCCUCGACUGGAAG	16,692
gga-miR-18a	UAAGGUGCAUCUAGUGCAGAUA	15,146
gga-miR-199*	UACAGUAGUCUGCACAUUGG	14,169
gga-miR-16	UAGCAGCACGUAAAUAUUGGUG	12,287

FIG 1 The E (XSR) element in the ALV-J genome encodes a novel microRNA. (A) Genomic location of E (XSR) miRNA. The structure of the ALV-J provirus HPRS-103 genome with the long terminal repeat (LTR), *gag/pol*, *env* (gp85 and gp37), redundant transmembrane (rTM), and the E element is shown. The secondary structure of E (XSR) pre-miRNA predicted using the MFOLD algorithm is shown on the top. The mature miRNA strands are indicated in red. The mature miRNAs with the number of reads for each strand recovered from deep sequencing are indicated. (B) Reads representing the major populations of E (XSR) miRNA-5p and E (XSR) miRNA-3p aligned against the precursor sequence. Dot-bracket notation was produced by applying RNAfold. (C) Most-often-sequenced small RNA sequences from IAH30 that map to *Gallus gallus* miRNAs (gga prefix) listed in miRBase.

dates. Closer examination of the sequence revealed that it is derived from the hairpin structure of the E (XSR) element (Fig. 1A). During miRNA biogenesis, processing of the pre-miRNA yields the more abundant miRNA and less abundant passenger strands (33). From our data, we conclude that the abundant species that mapped to 5′ arm of the stem-loop sequence is likely to be the miRNA strand, while the less abundant sequence that mapped to 3′ arm of the stem-loop represented the passenger strand. Importantly, the 5′ arm of the stem-loop sequence is the most abundant species present in the IAH30 library (Fig. 1C). Taken together, these results strongly suggested that we had identified a pre-miRNA encoded within the E (XSR) element of ALV-J.

In order to confirm that these novel miRNAs identified in the IAH30 cell line were indeed derived from the E (XSR) sequences in the HPRS-103 genome and did not originate from the turkey cells, we carried out Northern blotting hybridization with the miRNA strand probe. Northern blotting detected both mature miRNA and pre-miRNA in IAH30 cells. No miRNAs were detected with RNA extracted from the reticuloendotheliosis virus T (REV-T)-transformed turkey cell line AVOL-1T and uninfected turkey spleen cells (TSCs) (Fig. 2A). This finding was further confirmed by TaqMan miRNA assays (Fig. 2B). IAH30 is a turkey macrophage cell line transformed by strain 966, an acutely transforming virus derived from HPRS-103. To test that the expression of the E (XSR)-derived miRNA is not limited to the transformed macrophages of turkey origin, we also examined the primary chicken macrophages infected with either HPRS-103 or 966 virus. RNA isolated from the infected cells was tested for E (XSR) miRNA

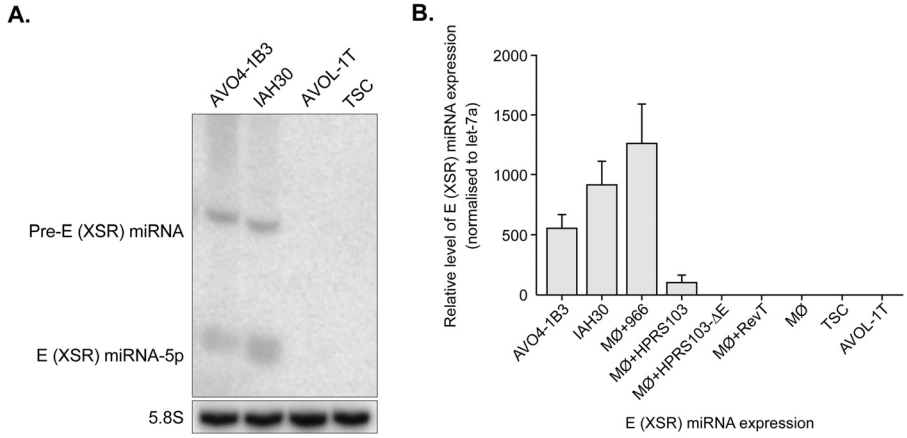


FIG 2 E (XSR) miRNA is expressed in ALV-J-transformed cells. (A) Northern blot analysis of RNA from the blastoderm cell line AVO4-1B3, turkey macrophage cell line IAH30, REV-T-transformed turkey spleen cell line AVOL-1T, and turkey spleen cells (TSCs) demonstrating the expression of mature and pre-miRNA of E (XSR) miRNA. The cellular 5.8S RNA served as the loading control. (B) E (XSR) miRNA expression levels determined by qRT-PCR. Relative expression of E (XSR) miRNA measured in RNA extracted from AVO4-1B3, IAH30, and chicken macrophages infected with either 966 (MΦ+966), HPRS-103 (MΦ+HPRS103), HPRS-103 with a deletion of the E element (MΦ+HPRS103-ΔE), or REV-T (MΦ+RevT) as well as AVOL-1T cells and TSCs compared to uninfected macrophage (MΦ). Results represent the means of triplicate assays with error bars showing the standard error of the means.

expression using a TaqMan assay. Indeed, E (XSR) miRNA was detected in cells infected with both types of viruses (Fig. 2B). Furthermore, E (XSR) miRNA was also detected in AVO4-1B3 cells, a chicken blastoderm cell line transformed by another acutely transforming ALV-J isolate, 1B (27), by both Northern blotting and TaqMan miRNA assay (Fig. 2A and B). The E (XSR) miRNA was not detected with RNA extracted from macrophages infected with HPRS-103 with a deletion of E (XSR) (7) or from uninfected or REV-T-infected macrophages (Fig. 2 B). The fact that E (XSR) miRNA was expressed in different cell types infected with a number of different ALV-J virus strains further confirmed that E (XSR) miRNA is indeed a genuine miRNA encoded by ALV-J virus.

E (XSR) miRNA sequence is highly conserved. We next examined the evolutionary conservation of E (XSR) miRNA by aligning all pre-miRNA sequences of NCBI-deposited ALV subgroup A, subgroup J, and Rous sarcoma virus isolates with E element sequences along with the pre-miRNA sequence of E (XSR) miRNA from HPRS-103 (see Fig. S1 in the supplemental material). Within the 64 aligned pre-miRNA sequences of the E (XSR) element, the miRNA strand (5′ arm, miRNA-5p) was identical in 46 isolates, with 16 isolates showing a single nucleotide change and the remaining 2 isolates showing 2-nucleotide differences. The seed region showed only a single nucleotide substitution among all the sequences, suggesting evolutionary pressure on maintaining the sequence of this region, potentially for modulating the expression of its targets. The sequence of the passenger strand (3′ arm, miRNA-3p), on the other hand, showed more substitutions. The loop region was highly conserved, and despite deletions or insertions in this region, all of the pre-miRNA sequences were able to form hairpin structures (data not shown). The fact that the miRNA sequence is well conserved across all known ALV-J isolates with an E (XSR) element suggests that this newly identified ALV-J miRNA may have a conserved functional role.

E (XSR) miRNA is processed by the canonical miRNA biogenesis pathway. The vast majority of viral miRNAs are transcribed by RNA polymerase II (Pol II) before being processed by RNase III enzymes Drosha and Dicer. The exceptions are the Pol III-derived tRNA-like precursor structures of mouse hepatitis virus 68 (MHV68) (32, 34) and the recently reported BLV miRNAs (22, 23, 35). In the absence of detecting any sequence motifs of Pol III promoter elements (22, 23) in the E (XSR) flanking sequence, we hypothesized that this miRNA is Pol II derived. To test this, we cloned the stem-loop structure sequence together with approximately 150 nucleotides of flanking sequence downstream of the cytomegalovirus (CMV) promoter into pcDNA3 vector (Invitrogen). High-level expression of E (XSR) miRNA detected by Northern blotting in transfected HEK293T cells suggested that the E (XSR) miRNA is processed by the Pol II promoter (Fig. 3A). The hypothesis of Pol II driving expression was further explored by testing the blockage of miRNA production by treatment with α-amanitin, a selective inhibitor of Pol II (22, 32). We transfected HEK293T cells with the plasmid expressing E (XSR) miRNA in pcDNA3 or pGEM-T easy vector in the presence or absence of α-amanitin. As shown in Fig. 3A, α-amanitin inhibits both pre-miRNA and mature miRNA of the E (XSR) miRNA, thus supporting our initial Pol II prediction. A cellular miRNA miR-155 transcribed under the control of the Pol II promoter pEF6-Bic and a virus miRNA, MHV68-miR-M1-7-3p (kindly provided by C. Sullivan), transcribed under the control of a Pol III promoter were used as controls. As expected, expression of both E (XSR) miRNA and miR-155 was blocked by α-amanitin treatment, whereas the MHV68-miR-M1-7-3p miRNA was resistant (Fig. 3A). This further confirms that the expression of E (XSR) miRNA is driven by Pol II promoter but not Pol III promoter.

Drosha is a key enzyme in microRNA biogenesis, generating the pre-miRNA by excising most pre-miRNA structures from Pol II transcripts (36). To determine whether Drosha contributes to E (XSR) miRNA processing, we used RNA interference (RNAi) to knock down Drosha in HEK293T cells. Two Silencer Select validated siRNAs against human Drosha (Life Technologies) were used for the knockdown of Drosha expression. We cotransfected HEK293T cells with siRNA against Drosha and vectors expressing either E (XSR) miRNA or control miRNA miR-155 (Fig. 3B). Knockdown of Drosha by either siRNA resulted in a marked de-

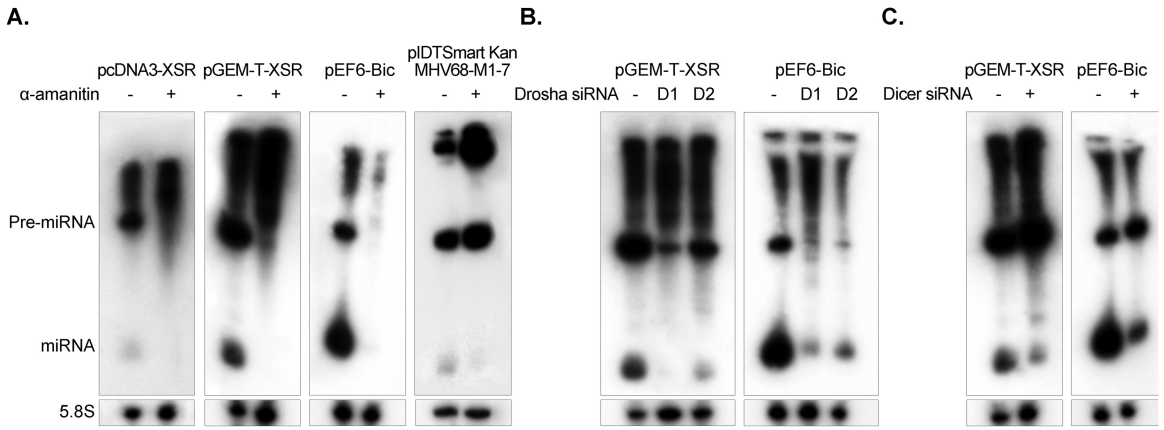


FIG 3 E (XSR) miRNA is processed by the canonical miRNA biogenesis pathway. (A) Northern blot analysis of E (XSR) miRNA expression plasmids pcDNA3-XSR and pGEM-T-XSR, plasmid pEF6-Bic expressing *Gallus gallus* miR-155 (gga-miR-155), and MHV68-miR-M1-7 expression plasmid pIDTSmart-Kan-MHV68-miR-M1-7 transfected into HEK293T cells with or without treatment of the Pol II inhibitor α-amanitin. The cellular 5.8S RNA was used as a loading control. (B) Northern blot analysis of E (XSR) miRNA and gga-miR-155 expression vectors transfected into HEK293T cells with or without siRNA against human Drosha (D1 and D2). 5.8S RNA was used as a loading control. (C) Northern blot analysis of E (XSR) miRNA and gga-miR-155 expression vectors transfected into HEK293T cells with or without siRNA against human Dicer. 5.8S RNA was used as a loading control.

crease in the miR-155 expression which is known to be Drosha dependent (Fig. 3B). Similar results were obtained with E (XSR) miRNA expression following Drosha knockdown, suggesting that E (XSR) miRNA expression is also Drosha dependent.

Dicer is the second RNA III enzyme in the miRNA biogenesis pathway responsible for cleavage of pre-miRNA to generate mature miRNA. To confirm the contribution of Dicer to E (XSR) miRNA expression, we cotransfected HEK293T cells with validated siRNA against Dicer (Life Technologies) and vectors expressing either E (XSR) miRNA or control miRNA miR-155 (Fig. 3C). As expected, knockdown of Dicer resulted in an increase of the ratio of pre-miRNA to miRNA for both E (XSR) miRNA and Dicer-dependent miR-155 (Fig. 3C). Thus, we conclude that E (XSR) miRNA is processed by the canonical miRNA biogenesis pathway.

E (XSR) miRNA is a biologically functional miRNA. To determine whether the E (XSR) miRNA is biologically functional, psiCHECK-2 luciferase-based reporter plasmids, bearing four tandem repeats of artificial target sites of perfect complementarity to the miRNA sequence inserted into the 3′ UTR (Fig. 4B), either were cotransfected into DF-1 cells along with the E (XSR) miRNA expression plasmid or along with the E (XSR) miRNA mimics (Qiagen) into either DF-1 cells and Vero cells using Lipofectamine 2000 (Invitrogen) or chicken embryo fibroblasts (CEF) using Lipofectamine (Invitrogen), or they were transfected into IAH30 using Nucleofactor Transfection Kit T (Lonza). The expression of E (XSR) miRNA from transfected DF-1 and in IAH30 cells was confirmed by quantitative TaqMan miRNA assay (Fig. 4A). As shown in Fig. 4B, E (XSR) miRNA expressed from an expression plasmid was able to inhibit luciferase reporter expression by 56% in DF-1 cells relative to expression from a mutant reporter plasmid with four mutated nucleotides in the seed region. E (XSR) miRNA mimics reduced the luciferase level by 68% in DF-1 cells, 88% in CEFs, and 98% in Vero cells, and the endogenous E (XSR) miRNA in IAH30 cells could inhibit luciferase expression by 65%. Thus, the reporter assay demonstrated that the miRNA is active within the RNA-induced silencing complex (RISC) and that the E (XSR) element is processed into functional mature miRNA in DF-1 cells.

DISCUSSION

Of the 295 virus-encoded miRNAs deposited in the miRBase database, the vast majority are encoded by DNA viruses. The small size and lack of immunogenicity, combined with the ability for specific repression of the expression of multiple target transcripts, make the miRNAs ideal tools for the viruses to reshape gene expression in an infected cell to favor viral replication and pathogenesis. Although there is a long way to go to gain significant understanding of how these miRNAs function and of the portfolio of their targets, it is clear that these small but effective regulators of gene expression play a key role in virus biology.

Among all viruses, the members of the family *Herpesviridae* account for the majority of the currently known virus-encoded miRNAs (19, 37). In addition to the herpesviruses, a small number of other nuclear DNA viruses, particularly polyomaviruses, have also been shown to encode miRNAs or miRNA-like molecules (38–45). Furthermore, no viral miRNAs have been identified using low- or high-throughput sequencing of RNA from cultured cells infected with any of several different RNA viruses, including hepatitis C virus (HCV), yellow fever virus (YFV), West Nile virus (WNV), human papillomavirus (HPV), vesicular stomatitis virus (VSV), dengue virus, polio virus, human T-cell leukemia virus type 1 (HTLV-1), and influenza A virus (21, 34, 46–48). Although there have been several reports suggesting that HIV, an RNA virus whose genome is reverse transcribed and incorporated into the host DNA, may also encode miRNAs, these reports are controversial (21, 34). Thus, a widely accepted example of any naturally RNA virus-encoded miRNA was lacking until the recent report of miRNAs encoded by bovine leukemia virus (BLV), a retrovirus with an RNA genome (22, 23). The most compelling hypothesis for the lack of miRNA sequences in the RNA virus genome is that excision of an miRNA from an RNA virus would result in the cleavage and ultimately the destruction of the viral genomic RNA, which would likely inhibit virus replication. This hypothesis is

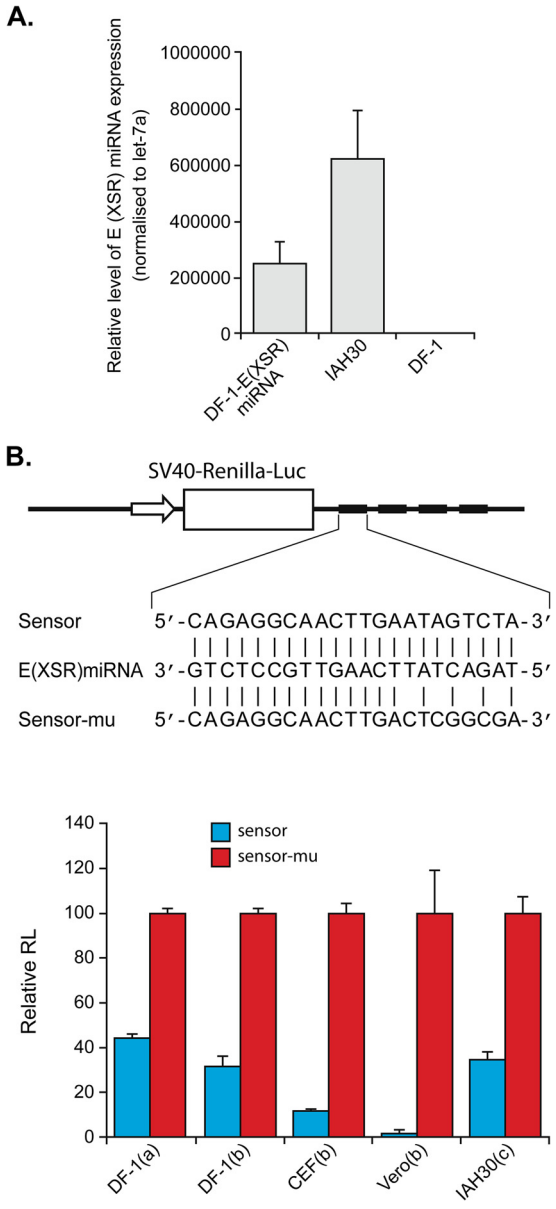


FIG 4 E (XSR) miRNA is a biologically functional miRNA. (A) E (XSR) miRNA expression levels determined by qRT-PCR. Relative expression of E (XSR) miRNA measured in RNA extracted from miRNA expression plasmid-transfected DF-1 and IAH30 cells compared to untransfected DF-1 cells. Results represent the means of triplicate assays, with error bars showing the standard errors of the means. (B) The top panel shows a sensor construct with four tandem repeats of artificial target sites of perfect complementarity to the miRNA sequence inserted into the 3′ UTR at the end of the *Renilla* luciferase gene in psiCHECK-2 vector (sensor) and a sensor mutant (sensor-mu) construct with four mutated nucleotides in the seed region. The bottom panel shows repression of luciferase by the sensor construct relative to the mutant construct after cotransfection with an E (XSR) miRNA expression plasmid into DF-1 cells (a) or cotransfection with E (XSR) miRNA mimics into DF-1, CEF, and Vero cells (b) and transfection into IAH30 (c). The relative expression of *Renilla* luciferase (RL) was determined with the normalized levels of firefly luciferase. For each sample, values from four replicates representative of at least two independent experiments were used in the analysis. SV40, simian virus 40.

supported by the finding that BLV can overcome this obstacle by encoding pre-miRNA structures that are only competent to be processed into miRNAs when they are generated from subgenomic Pol III-derived transcripts (22, 23).

In spite of this theoretical barrier, here we provide evidence that ALV-J uses Pol II for the production of high-level E (XSR) miRNA. Subsequently, we have shown that the processing of the E (XSR) miRNA is Drosha and Dicer dependent. Taken together, this is the first example of an RNA virus that encodes an miRNA using the canonical miRNA biogenesis pathway. This suggests that the RNA viruses could utilize different strategies to express their own miRNAs and that they could tolerate *cis* cleavage within the genome during pre-miRNA processing. Indeed, retroviruses, a flavivirus, and influenza virus can be engineered to express biologically active miRNAs or miRNA-like molecules (49–51). The evidence of BLV miRNAs transcribed by the Pol III promoter and ALV E (XSR) miRNA transcribed by the Pol II promoter suggests that future miRNA discovery efforts could be directed to other retroviruses.

One of the conspicuous findings from the analysis of the miRNA sequences of the IAH30 library was that E (XSR) miRNA accounted for a quarter of the 1.469×10^6 sequences of the IAH30 miRNAome. An increased proportion of virus-encoded miRNAs to host-encoded miRNAs is not uncommon in transformed cell lines. For example, miRNAs encoded by Kaposi's sarcoma-associated herpesvirus and Epstein-Barr virus (EBV) accounted for ~40% of the entire miRNA pool identified from the BC-1 cell line coinfected with these two viruses (52). The total proportion of virus-encoded miRNAs of Marek's disease virus type 1 (MDV-1) and Marek's disease virus type 2 (MDV-2) in an MSB-1 cell library was 61% (53). The high-level expression of viral miRNAs has been linked to their role in virus-induced oncogenesis since the cluster 1 miRNAs of MDV-1 and mdv1-mir-M4-5p, a member of cluster 1 miRNA and a functional ortholog of gga-mir-155 which are highly expressed in the transformed cell lines and in tumors, have been shown to play a key role in MDV-1-induced tumorigenesis (54). A high level of expression of E (XSR) miRNA in the IAH30 cell line suggests that this miRNA has a major role in ALV-J pathogenesis and neoplastic transformation. Although the E element *per se* is not absolutely essential for tumor induction by this subgroup of viruses, our previous work comparing the oncogenicity of viruses derived from the parental HPRS-103 virus and from HPRS-103 with a deletion of the E element in two genetically distinct lines of birds showed that the E element does contribute to oncogenicity in certain genetic lines of chicken (7). Future studies comparing the genomes of these lines could provide insights into the polymorphisms, including those in any potential E (XSR) miRNA target sites that could account for such differential susceptibility phenotypes among these lines. The discoveries of the virus-encoded miRNA targets would help us to get a clearer understanding of the role played by the viral miRNAs.

In summary, we demonstrated that an RNA virus expresses abundant, evolutionarily conserved miRNA using the canonical miRNA biogenesis pathway. The identification of this novel potentially functional miRNA adds yet another regulatory mechanism in the pathobiology of ALV and RSV.

ACKNOWLEDGMENT

This work was supported by the Biotechnology and Biological Sciences Research Council, United Kingdom.

REFERENCES

1. **Payne LN.** 1998. Retrovirus-induced disease in poultry. Poult. Sci. **77:** 1204–1212.
2. **Ruddell A.** 1995. Transcription regulatory elements of the avian retroviral long terminal repeat. Virology **206:**1–7. http://dx.doi.org/10.1016/S0042-6822(95)80013-1.
3. **Banks JD, Beemon KL, Linial ML.** 1997. RNA regulatory elements in the genomes of simple retroviruses. Semin. Virol. **8:**194–204. http://dx.doi.org/10.1006/smvy.1997.0122.
4. **Bizub D, Katz RA, Skalka AM.** 1984. Nucleotide sequence of noncoding regions in Rous-associated virus-2: comparisons delineate conserved regions important in replication and oncogenesis. J. Virol. **49:**557–565.
5. **Schwartz DE, Tizard R, Gilbert W.** 1983. Nucleotide sequence of Rous sarcoma virus. Cell **32:**853–869. http://dx.doi.org/10.1016/0092-8674(83)90071-5.
6. **Tsichlis PN, Donehower L, Hager G, Zeller N, Malavarca R, Astrin S, Skalka AM.** 1982. Sequence comparison in the crossover region of an oncogenic avian retrovirus recombinant and its nononcogenic parent: genetic regions that control growth rate and oncogenic potential. Mol. Cell. Biol. **2:**1331–1338.
7. **Chesters PM, Smith LP, Nair V.** 2006. E (XSR) element contributes to the oncogenicity of avian leukosis virus (subgroup J). J. Gen. Virol. **87:**2685–2692. http://dx.doi.org/10.1099/vir.0.81884-0.
8. **Bai J, Payne LN, Skinner MA.** 1995. HPRS-103 (exogenous avian leukosis virus, subgroup J) has an env gene related to those of endogenous elements EAV-0 and E51 and an E element found previously only in sarcoma viruses. J. Virol. **69:**779–784.
9. **Cui Z, Du Y, Zhang Z, Silva RF.** 2003. Comparison of Chinese field strains of avian leukosis subgroup J viruses with prototype strain HPRS-103 and United States strains. Avian Dis. **47:**1321–1330. http://dx.doi.org/10.1637/6085.
10. **Liu C, Zheng S, Wang Y, Jing L, Gao H, Gao Y, Qi X, Qin L, Pan W, Wang X.** 2011. Detection and molecular characterization of recombinant avian leukosis viruses in commercial egg-type chickens in China. Avian Pathol. **40:**269–275. http://dx.doi.org/10.1080/03079457.2011.560932.
11. **Lupiani B, Williams SM, Silva RF, Hunt HD, Fadly AM.** 2003. Pathogenicity of two recombinant avian leukosis viruses. Avian Dis. **47:**425–432. http://dx.doi.org/10.1637/0005-2086(2003)047%5B0425%3APOTRAL%5D2.0.CO%3B2.
12. **Payne LN, Gillespie AM, Howes K.** 1992. Myeloid leukaemogenicity and transmission of the HPRS-103 strain of avian leukosis virus. Leukemia **6:**1167–1176.
13. **Wu X, Qian K, Qin A, Shen H, Wang P, Jin W, Eltahir YM.** 2010. Recombinant avian leukosis viruses of subgroup J. isolated from field infected commercial layer chickens with hemangioma and myeloid leukosis possess an insertion in the E element. Vet. Res. Commun. **34:**619–632. http://dx.doi.org/10.1007/s11259-010-9436-8.
14. **Hue D, Dambrine G, Denesvre C, Laurent S, Wyers M, Rasschaert D.** 2006. Major rearrangements in the E element and minor variations in the U3 sequences of the avian leukosis subgroup J provirus isolated from field myelocytomatosis. Arch. Virol. **151:**2431–2446. http://dx.doi.org/10.1007/s00705-006-0811-2.
15. **Pan W, Gao Y, Sun F, Qin L, Liu Z, Yun B, Wang Y, Qi X, Gao H, Wang X.** 2011. Novel sequences of subgroup J avian leukosis viruses associated with hemangioma in Chinese layer hens. Virol. J. **8:**552. http://dx.doi.org/10.1186/1743-422X-8-552.
16. **Zavala G, Cheng S, Jackwood MW.** 2007. Molecular epidemiology of avian leukosis virus subgroup J. and evolutionary history of its 3′ untranslated region. Avian Dis. **51:**942–953. http://dx.doi.org/10.1637/0005-2086(2007)51[942:MEOALV]2.0.CO;2.
17. **Shi M, Tian M, Liu C, Zhao Y, Lin Y, Zou N, Liu P, Huang Y.** 2011. Sequence analysis for the complete proviral genome of subgroup J avian leukosis virus associated with hemangioma: a special 11 bp deletion was observed in U3 region of 3′UTR. Virol. J. **8:**158. http://dx.doi.org/10.1186/1743-422X-8-158.
18. **Griffiths-Jones S.** 2010. miRBase: microRNA sequences and annotation. Curr. Protoc. Bioinformatics Chapter 12:Unit 12.9. http://dx.doi.org/10.1002/0471250953.bi1209s29.
19. **Cullen BR.** 2009. Viral and cellular messenger RNA targets of viral microRNAs. Nature **457:**421–425. http://dx.doi.org/10.1038/nature07757.
20. **Kim do, N, Lee SK.** 2012. Biogenesis of Epstein-Barr virus microRNAs. Mol. Cell. Biochem. **365:**203–210. http://dx.doi.org/10.1007/s11010-012-1261-7.
21. **Lin J, Cullen BR.** 2007. Analysis of the interaction of primate retroviruses with the human RNA interference machinery. J. Virol. **81:**12218–12226. http://dx.doi.org/10.1128/JVI.01390-07.
22. **Kincaid RP, Burke JM, Sullivan CS.** 2012. RNA virus microRNA that mimics a B-cell oncomiR. Proc. Natl. Acad. Sci. U. S. A. **109:**3077–3082. http://dx.doi.org/10.1073/pnas.1116107109.
23. **Rosewick N, Momont M, Durkin K, Takeda H, Caiment F, Cleuter Y, Vernin C, Mortreux F, Wattel E, Burny A, Georges M, Van den Broeke A.** 2013. Deep sequencing reveals abundant noncanonical retroviral microRNAs in B-cell leukemia/lymphoma. Proc. Natl. Acad. Sci. U. S. A. **110:**2306–2311. http://dx.doi.org/10.1073/pnas.1213842110.
24. **Himly M, Foster DN, Bottoli I, Iacovoni JS, Vogt PK.** 1998. The DF-1 chicken fibroblast cell line: transformation induced by diverse oncogenes and cell death resulting from infection by avian leukosis viruses. Virology **248:**295–304. http://dx.doi.org/10.1006/viro.1998.9290.
25. **Lawson S, Rothwell L, Lambrecht B, Howes K, Venugopal K, Kaiser P.** 2001. Turkey and chicken interferon-gamma, which share high sequence identity, are biologically cross-reactive. Dev. Comp. Immunol. **25:**69–82. http://dx.doi.org/10.1016/S0145-305X(00)00044-6.
26. **Chesters PM, Howes K, McKay JC, Payne LN, Venugopal K.** 2001. Acutely transforming avian leukosis virus subgroup J. strain 966: defective genome encodes a 72-kilodalton Gag-Myc fusion protein. J. Virol. **75:**4219–4225. http://dx.doi.org/10.1128/JVI.75.9.4219-4225.2001.
27. **Venugopal K, Howes K, Flannery DM, Payne LN.** 2000. Isolation of acutely transforming subgroup J. avian leukosis viruses that induce erythroblastosis and myelocytomatosis. Avian Pathol. **29:**497–503. http://dx.doi.org/10.1080/030794500750047252.
28. **Zhao Y, Yao Y, Xu H, Lambeth L, Smith LP, Kgosana L, Wang X, Nair V.** 2009. A functional MicroRNA-155 ortholog encoded by the oncogenic Marek's disease virus. J. Virol. **83:**489–492. http://dx.doi.org/10.1128/JVI.01166-08.
29. **Yao Y, Smith LP, Petherbridge L, Watson M, Nair V.** 2012. Novel microRNAs encoded by duck enteritis virus. J. Gen. Virol. **93:**1530–1536. http://dx.doi.org/10.1099/vir.0.040634-0.
30. **Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods **5:**621–628. http://dx.doi.org/10.1038/nmeth.1226.
31. **Yao Y, Zhao Y, Xu H, Smith LP, Lawrie CH, Sewer A, Zavolan M, Nair V.** 2007. Marek's disease virus type 2 (MDV-2)-encoded microRNAs show no sequence conservation with those encoded by MDV-1. J. Virol. **81:**7164–7170. http://dx.doi.org/10.1128/JVI.00112-07.
32. **Bogerd HP, Karnowski HW, Cai X, Shin J, Pohlers M, Cullen BR.** 2010. A mammalian herpesvirus uses noncanonical expression and processing mechanisms to generate viral microRNAs. Mol. Cell **37:**135–142. http://dx.doi.org/10.1016/j.molcel.2009.12.016.
33. **Kim VN.** 2005. Small RNAs: classification, biogenesis, and function. Mol. Cells **19:**1–15.
34. **Pfeffer S, Sewer A, Lagos-Quintana M, Sheridan R, Sander C, Grasser FA, van Dyk LF, Ho CK, Shuman S, Chien M, Russo JJ, Ju J, Randall G, Lindenbach BD, Rice CM, Simon V, Ho DD, Zavolan M, Tuschl T.** 2005. Identification of microRNAs of the herpesvirus family. Nat. Methods **2:**269–276. http://dx.doi.org/10.1038/nmeth746.
35. **Cullen BR.** 2012. MicroRNA expression by an oncogenic retrovirus. Proc. Natl. Acad. Sci. U. S. A. **109:**2695–2696. http://dx.doi.org/10.1073/pnas.1200328109.
36. **Bartel DP.** 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell **116:**281–297. http://dx.doi.org/10.1016/S0092-8674(04)00045-5.
37. **Gottwein E, Cullen BR.** 2008. Viral and cellular microRNAs as determinants of viral pathogenesis and immunity. Cell Host Microbe **3:**375–387. http://dx.doi.org/10.1016/j.chom.2008.05.002.
38. **Aparicio O, Razquin N, Zaratiegui M, Narvaiza I, Fortes P.** 2006. Adenovirus virus-associated RNA is processed to functional interfering RNAs involved in virus production. J. Virol. **80:**1376–1384. http://dx.doi.org/10.1128/JVI.80.3.1376-1384.2006.
39. **Cantalupo P, Doering A, Sullivan CS, Pal A, Peden KW, Lewis AM, Pipas JM.** 2005. Complete nucleotide sequence of polyomavirus SA12. J. Virol. **79:**13094–13104. http://dx.doi.org/10.1128/JVI.79.20.13094-13104.2005.
40. **Hussain M, Taft RJ, Asgari S.** 2008. An insect virus-encoded microRNA regulates viral replication. J. Virol. **82:**9164–9170. http://dx.doi.org/10.1128/JVI.01109-08.
41. **Seo GJ, Chen CJ, Sullivan CS.** 2009. Merkel cell polyomavirus encodes a microRNA with the ability to autoregulate viral gene expression. Virology **383:**183–187. http://dx.doi.org/10.1016/j.virol.2008.11.001.
42. **Seo GJ, Fink LH, O'Hara B, Atwood WJ, Sullivan CS.** 2008. Evolution-

arily conserved function of a viral microRNA. J. Virol. **82:**9823–9828. http://dx.doi.org/10.1128/JVI.01144-08.
43. **Singh J, Singh CP, Bhavani A, Nagaraju J.** 2010. Discovering microRNAs from *Bombyx mori* nucleopolyhedrosis virus. Virology **407:**120–128. http://dx.doi.org/10.1016/j.virol.2010.07.033.
44. **Sullivan CS, Grundhoff AT, Tevethia S, Pipas JM, Ganem D.** 2005. SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. Nature **435:**682–686. http://dx.doi.org/10.1038/nature03576.
45. **Sullivan CS, Sung CK, Pack CD, Grundhoff A, Lukacher AE, Benjamin TL, Ganem D.** 2009. Murine polyomavirus encodes a microRNA that cleaves early RNA transcripts but is not essential for experimental infection. Virology **387:**157–167. http://dx.doi.org/10.1016/j.virol.2009.02.017.
46. **Parameswaran P, Sklan E, Wilkins C, Burgon T, Samuel MA, Lu R, Ansel KM, Heissmeyer V, Einav S, Jackson W, Doukas T, Paranjape S, Polacek C, dos Santos FB, Jalili R, Babrzadeh F, Gharizadeh B, Grimm D, Kay M, Koike S, Sarnow P, Ronaghi M, Ding SW, Harris E, Chow M, Diamond MS, Kirkegaard K, Glenn JS, Fire AZ.** 2010. Six RNA viruses and forty-one hosts: viral small RNAs and modulation of small RNA repertoires in vertebrate and invertebrate systems. PLoS Pathog. **6:**e1000764. http://dx.doi.org/10.1371/journal.ppat.1000764.
47. **Cai X, Li G, Laimins LA, Cullen BR.** 2006. Human papillomavirus genotype 31 does not express detectable microRNA levels during latent or productive virus replication. J. Virol. **80:**10890–10893. http://dx.doi.org/10.1128/JVI.01175-06.
48. **Umbach JL, Yen HL, Poon LL, Cullen BR.** 2010. Influenza A virus expresses high levels of an unusual class of small viral leader RNAs in infected cells. mBio **1**(4):e00204–10. http://dx.doi.org/10.1128/mBio.00204-10.
49. **Rouha H, Thurner C, Mandl CW.** 2010. Functional microRNA generated from a cytoplasmic RNA virus. Nucleic Acids Res. **38:**8328–8337. http://dx.doi.org/10.1093/nar/gkq681.
50. **Shapiro JS, Varble A, Pham AM, Tenoever BR.** 2010. Noncanonical cytoplasmic processing of viral microRNAs. RNA **16:**2068–2074. http://dx.doi.org/10.1261/rna.2303610.
51. **Varble A, Chua MA, Perez JT, Manicassamy B, Garcia-Sastre A, ten Oever BR.** 2010. Engineered RNA viral synthesis of microRNAs. Proc. Natl. Acad. Sci. U. S. A. **107:**11519–11524. http://dx.doi.org/10.1073/pnas.1003115107.
52. **Cai X, Lu S, Zhang Z, Gonzalez CM, Damania B, Cullen BR.** 2005. Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. Proc. Natl. Acad. Sci. U. S. A. **102:**5570–5575. http://dx.doi.org/10.1073/pnas.0408192102.
53. **Yao Y, Zhao Y, Xu H, Smith LP, Lawrie CH, Watson M, Nair V.** 2008. MicroRNA profile of Marek's disease virus-transformed T-cell line MSB-1: predominance of virus-encoded microRNAs. J. Virol. **82:**4007–4015. http://dx.doi.org/10.1128/JVI.02659-07.
54. **Zhao Y, Xu H, Yao Y, Smith LP, Kgosana L, Green J, Petherbridge L, Baigent SJ, Nair V.** 2011. Critical role of the virus-encoded microRNA-155 ortholog in the induction of Marek's disease lymphomas. PLoS Pathog. **7:**e1001305. http://dx.doi.org/10.1371/journal.ppat.1001305.

Journal of Virology

RNA Interference Targets Arbovirus Replication in Culicoides Cells

Esther Schnettler, Maxime Ratinier, Mick Watson, Andrew E. Shaw, Melanie McFarlane, Mariana Varela, Richard M. Elliott, Massimo Palmarini and Alain Kohl

J. Virol. **2013, 87(5):2441. DOI: 10.1128/JVI.02848-12.**
Published Ahead of Print 26 December 2012.

JVI Journals.ASM.org

RNA Interference Targets Arbovirus Replication in *Culicoides* Cells

Esther Schnettler,[a] Maxime Ratinier,[a] Mick Watson,[b] Andrew E. Shaw,[a] Melanie McFarlane,[a] Mariana Varela,[a] Richard M. Elliott,[c] Massimo Palmarini,[a] Alain Kohl[a]

MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom[a]; The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, United Kingdom[b]; Biomedical Sciences Research Complex, School of Biology, University of St. Andrews, North Haugh, St. Andrews, United Kingdom[c]

Arboviruses are transmitted to vertebrate hosts by biting arthropod vectors such as mosquitoes, ticks, and midges. These viruses replicate in both arthropods and vertebrates and are thus exposed to different antiviral responses in these organisms. RNA interference (RNAi) is a sequence-specific RNA degradation mechanism that has been shown to play a major role in the antiviral response against arboviruses in mosquitoes. *Culicoides* midges are important vectors of arboviruses, known to transmit pathogens of humans and livestock such as bluetongue virus (BTV) (*Reoviridae*), Oropouche virus (*Bunyaviridae*), and likely the recently discovered Schmallenberg virus (*Bunyaviridae*). In this study, we investigated whether *Culicoides* cells possess an antiviral RNAi response and whether this is effective against arboviruses, including those with double-stranded RNA (dsRNA) genomes, such as BTV. Using reporter gene-based assays, we established the presence of a functional RNAi response in *Culicoides sonorensis*-derived KC cells which is effective in inhibiting BTV infection. Sequencing of small RNAs from KC and *Aedes aegypti*-derived Aag2 cells infected with BTV or the unrelated Schmallenberg virus resulted in the production of virus-derived small interfering RNAs (viRNAs) of 21 nucleotides, similar to the viRNAs produced during arbovirus infections of mosquitoes. In addition, viRNA profiles strongly suggest that the BTV dsRNA genome is accessible to a Dicer-type nuclease. Thus, we show for the first time that midge cells target arbovirus replication by mounting an antiviral RNAi response mainly resembling that of other insect vectors of arboviruses.

Biting arthropods such as mosquitoes, ticks, and midges can transmit a variety of viruses (arboviruses) belonging to the *Flaviviridae*, *Bunyaviridae*, *Togaviridae*, and *Reoviridae* families. Arboviruses actively replicate in both their arthropod vector and vertebrate host. At present, mosquito-borne viruses are probably the best-studied arboviruses. Among these are viruses of particular relevance to public health, including members of the *Flaviviridae* family, such as dengue virus (DENV), West Nile virus (WNV), and Japanese encephalitis virus (JEV), or alphaviruses of the *Togaviridae* family, such as chikungunya virus (CHIKV) (1).

Midge-borne viruses also impact on public health. Oropouche virus (OROV) infection can result in Oropouche fever, one of the most important arboviral diseases in America (mainly in the Amazon region, Panama, and Caribbean) (2, 3). *Culicoides* are biting haematophagous midges belonging to the family Ceratopogonidae. Importantly, 96% of the >1,400 identified species attack mammals, including humans. *Culicoides* are well-known vectors of protozoans, filarial worms, and viruses (3), and more than 50 viruses belonging to the *Bunyaviridae*, *Reoviridae*, and *Rhabdoviridae* families have been isolated from different *Culicoides* species. While some of these may be accidental infections, around 45% of isolated viruses are specific to *Culicoides* species, including those known to cause infections of livestock all over the world, such as African horse sickness virus (AHSV), bluetongue virus (BTV) (*Reoviridae*) (3), and the recently discovered Schmallenberg virus (SBV) (*Bunyaviridae*) (4).

As arboviruses require vectors for successful transmission between vertebrate hosts, there is evolutionary pressure on keeping the right balance between virus replication and vector survival. Recent research on mosquito-arbovirus interactions indicates that innate immune responses such as RNA interference (RNAi) are key factors in restricting arbovirus replication (5–12), as detailed in recent reviews (13, 14). Similar research on midge-transmitted arboviruses has not been carried out despite the fact that important arboviruses are transmitted by these vectors.

RNAi has been shown to be an important and possibly the major antiviral response in mosquitoes (13, 14). RNAi consists of different pathways that perform sequence-specific targeting of RNA. However, the exogenous small interfering RNA (siRNA) pathway is of particular interest given its antiviral function, as demonstrated in different organisms, including drosophila and mosquitoes (13–15). The mosquito exogenous RNAi pathway is induced by virus-derived long double-stranded RNA (dsRNA) either derived from replication intermediates or secondary structures that are targeted by the RNase III enzyme Dicer-2 (Dcr-2) and cut into virus-derived small interfering RNAs (viRNAs) of mainly 21 nucleotides (nt) in length, as is assumed through comparisons to *Drosophila melanogaster* (5, 7, 11, 12, 16–20). These viRNAs are taken up by the RNA-induced silencing complex (RISC), harboring an argonaute protein (Ago-2) as the catalytic compound. viRNAs are then unwound, and one strand is kept in the RISC to be used as a guide to find complementary viral RNA sequences. After base pairing, the catalytic domain of Ago-2 cleaves the target (viral) RNA, at least in the drosophila model, which silences viral infections (13, 14, 21–23). The exogenous siRNA pathway can

Received 11 October 2012 **Accepted** 17 December 2012
Published ahead of print 26 December 2012
Address correspondence to Alain Kohl, alain.kohl@glasgow.ac.uk.
E.S. and M.R. contributed equally to this work.

doi:10.1128/JVI.02848-12
The authors have paid a fee to allow immediate free access to this article.

also be artificially induced by the addition/transfection of long dsRNA or siRNA molecules, resulting in sequence-specific silencing. Key proteins of the RNA silencing pathways, such as Dcr-2 and Ago-2, have been shown to be conserved in drosophila and mosquitoes, and the effector mechanisms are likely to be similar. Other Dicer and Ago proteins are involved in a variety of small RNA silencing pathways, such as the microRNA pathway (13, 14, 23). A number of RNAi-competent mosquito cell lines, such as Aag2 (derived from *Aedes aegypti*) and U4.4 (derived from *Aedes albopictus*), as well as Dcr-2-deficient cell lines (C6/36 and C7-10, both derived from *Aedes albopictus*), have proven to be highly useful in studying mosquito RNAi responses (6, 10–12, 19, 24, 25). However, nothing is known about the presence and function of RNAi pathways, specifically the antiviral exogenous siRNA pathway, in midges and their derived cell lines.

BTV is one of the best-studied midge-borne viruses. It has been shown to replicate both in its arthropod vector and mammalian host (26, 27). BTV infection leads to persistent infection in infected adult *Culicoides* or midge-derived cell culture (28–30). This is in contrast to infected mammalian cells, which show strong cytopathic effects (30). Given the absence of studies on *Culicoides* RNAi pathways and antiviral mechanisms, nothing is known about the interactions of BTV with vector immune responses. Many *Culicoides* species have been identified as BTV vectors around the world, including *Culicoides imicola* in Africa (31) and Southern Europe (32), *Culicoides obsoletus* and *Culicoides pulicaris* in Central and Northern Europe (33, 34), and *Culicoides variipennis* and *Culicoides sonorensis* in America (35, 36). The BTV genome consists of 10 segments of dsRNA molecules (each comprising a coding and noncoding strand) that are packaged within a nonenveloped triple-layered icosahedral protein capsid (37–39) and direct the expression of 7 structural proteins (VP1 to VP7) and 4 distinct nonstructural proteins (NS1, NS2, NS3/NS3a, and NS4) (39–41). In contrast to the single-stranded RNA arboviruses with positive-sense (alphaviruses, flaviviruses) or negative-sense (bunyaviruses) RNA genomes that have been studied in mosquitoes or mosquito cell culture systems, the dsRNA nature of the BTV genome adds a layer of complexity for the antiviral RNAi response in insects. During the reovirus replication cycle, second-strand RNA synthesis is believed to occur only after assembly and consequently within the newly formed viral particles. As such, viral dsRNA is not necessarily accessible to the RNAi machinery. In addition to BTV, we are also investigating the RNAi response against SBV, an unrelated negative-strand RNA arbovirus. SBV is a recently emerged virus that affects ruminants causing mild disease (reduced milk production, pyrexia, and diarrhea) in adults and congenital malformations in stillborns or newborns (42, 43). SBV belongs to the genus *Orthobunyavirus* within the *Bunyaviridae* family and possesses a three-segmented negative-sense RNA genome. The large (L) segment encodes the RNA-dependent-RNA polymerase L, the medium (M) segment encodes a polyprotein that is cleaved into two glycoproteins (Gn and Gc) and a nonstructural protein (NSm), while the small (S) segment encodes the nucleoprotein (N) and a second nonstructural protein (NSs). SBV is believed to be transmitted by *Culicoides* species to susceptible mammalian hosts (4). Again, little is known about control of bunyavirus replication by RNAi responses of arthropod vectors. Recently, LaCrosse virus (LACV)-infected drosophila cells were shown to produce virus-specific 21-nt viRNAs mapping to all three viral segments, similar to what has been observed for other (mainly positive-strand RNA) arboviruses (6). In addition, LACV-infected *Aedes albopictus* C6/36 cells, known to be deficient in Dcr-2 activity, were shown to produce virus-specific small RNAs of different sizes (6, 44).

In this study, (i) we investigated the presence of a functional exogenous siRNA pathway in the *C. sonorensis*-derived KC cell line and (ii) we assessed whether this antiviral response targets midge-borne arboviruses with either dsRNA (BTV) or negative-strand RNA (SBV) genomes. We identified an exogenous siRNA pathway in KC cells that could be induced by dsRNA or viral infection and is effective against two arboviruses with highly different genome structures. A comparative analysis with BTV- or SBV-infected *Aedes aegypti* Aag2 cells suggests that the midge antiviral RNAi response against these viruses resembles mainly that of mosquitoes and points to conservation of key elements controlling arbovirus replication by RNAi in insect vectors.

MATERIALS AND METHODS

Cells. BSR cells, a clone of BHK-21 (kindly provided by Karl K. Conzelmann), were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 5% fetal bovine serum (FBS). BHK-21 cells were grown in Glasgow minimal essential medium (GMEM) supplemented with 10% newborn calf serum and 10% tryptose phosphate broth. CPT-Tert cells (45), immortalized sheep choroid plexus cells (kindly provided by D. Griffiths), were grown in Iscove's modified Dulbecco's medium (IMDM) supplemented with 10% FBS. Mammalian cell lines were cultured at 37°C in a 5% CO_2 humidified atmosphere. KC cells, obtained from *C. sonorensis* larvae, were grown in Schneider's insect medium supplemented with 10% FBS (46), *Aedes aegypti*-derived Aag2 mosquito cells were grown in L-15 medium supplemented with 10% FBS and 10% tryptose phosphate broth. Insect cells were maintained at 28°C.

Viruses and plasmids. BTV-1 was rescued by reverse genetics as previously described (41) and derived from the reference strain of BTV-1 originally isolated at the ARC-Onderstepoort Veterinary Institute. Virus stocks were prepared by infecting BSR cells at a low multiplicity of infection (MOI; 0.001) and harvesting the supernatant at 72 h postinfection. The virus suspension was centrifuged at 500 × *g* for 5 min. SBV (kindly provided by M. Beer) was initially isolated from blood of an infected cow and passaged once in KC cells and 6 times in BHK-21 cells. The virus was plaque purified, and stocks were produced in BHK-21 cells by infecting cells at a low MOI (0.01) and harvesting the supernatant at 120 h postinfection, followed by 20 min of centrifugation at 3,500 rpm. Virus titers of SBV and BTV-1 were established by standard plaque assays using CPT-Tert cells (47).

Expression vectors for invertebrate cells, pIZ-Fluc and pAcIE1-Rluc, expressing firefly (*FFluc*) and *Renilla* (*RLuc*) luciferases, respectively, have been previously described (48), and the fluorescein-labeled plasmid DNA was commercially obtained (Mirus).

Luciferase assays. Luciferase activities were determined using a dual-luciferase assay kit (Promega) on a GloMax-Multi+ microplate multimode reader following cell lysis in passive lysis buffer.

dsRNA production. dsRNA for the RNA silencing experiments was produced with the RNAi Megascript kit using gel-purified PCR products of a specific sequence (Table 1) flanked by T7 promoter sequences. Fluorescein-labeled dsRNA was produced by T7 RNA polymerase transcription (Invitrogen) on an enhanced green fluorescent protein (eGFP)-derived PCR product (using pC1-eGFP from Clontech as the template) with the fluorescein-labeled rNTP mix (Roche) by following the manufacturer's protocol. The DNA template and single-stranded RNA were removed by DNase1 and RNase A treatment (Ambion). dsRNA was then ethanol precipitated, dried, and resuspended in water.

TABLE 1 List of primer sequences used

Gene	Upstream/downstream primer sequence (5′→3′)[a]
BTV-1 NS1	*GTA ATA CGA CTC ACT ATA GGG* TCGGTGGG AATGGCTTAT
	GTA ATA CGA CTC ACT ATA GGG CTTTTCTG CATAGCATAGGGTG
eGFP, 400 nt	*GTA ATA CGA CTC ACT ATA GGG* GGCGTGC AGTGCTTCAGCCGC/
	GTA ATA CGA CTC ACT ATA GGG GTG GTTGTCGGGCAGCAGCAC
Firefly luciferase	*GTA ATA CGA CTC ACT ATA GGG* ATGGAAGC AGCCAAAAAC
	GTA ATA CGA CTC ACT ATA GGG TTACACG GCGATCTTTCC

[a] The T7 promoter region is indicated in italics.

Transfection and infection. In order to determine the transfection efficiency, 5×10^5 *C. sonorensis*-derived KC cells were seeded per well in 24-well plates with a glass bottom prior to transfection. DNA or dsRNA was incubated in the presence of a variety of transfection reagents (Fugene, Genejammer, and Lipofectamine 2000) and added to cells according to the manufacturer's protocol. In the case of Fugene and Genjammer, a ratio of 1 to 3 (micrograms of nucleic acid to transfection reagent) was used. At 24 h posttransfection, fluorescence in cells was analyzed using a Zeiss laser scanning microscopy (LSM) Meta microscope.

For reporter RNAi assays, 5×10^4 KC cells were seeded per well in 96-well plates and transfected with 250 ng pIZ-Fluc and 50 ng pAcIE1-Rluc using Fugene. Different concentrations of dsRNA, control (eGFP-specific) or targeting *FFluc*, were either cotransfected (5 ng, 1 ng, or 0.5 ng) in 50 μl of Schneider's medium, followed by the addition of 50 μl Schneider's medium with 30% fetal calf serum (FCS) at 2 h posttransfection, or added (300 ng or 100 ng) to cells in 100 μl Schneider's medium with 15% FCS at 24 h posttransfection. Luciferase activity was determined at 48 h posttransfection.

Assays to test dsRNA-induced antiviral activity were performed by seeding 1×10^6 KC cells per well in 24-well plates with glass bottoms. After 24 h, either 100 ng dsRNA, control (eGFP-specific) or specific for BTV NS1, was transfected using Lipofectamine 2000 or 500 ng dsRNA was added to the medium in the absence of a transfection reagent. KC cells were infected at an MOI of 0.2 with BTV-1 24 h posttransfection. At 2 h postinfection, the inocula were removed and cells were washed once with PBS. Supernatant was collected from cultured cells 24 h postinfection and centrifuged at $500 \times g$ for 5 min. Viral titers were subsequently determined by endpoint dilution analysis on BSR cells and expressed as $\log_{10}$ 50% tissue culture infective doses ($TCID_{50}$)/ml, calculated using the method of Reed and Muench (49). Cells were then fixed in 5% formaldehyde for 30 min at 24 h or 48 h postinfection (as indicated) and subsequently used for immunofluorescence assays or lysed for Western blot analysis.

***In vitro* growth kinetics of SBV.** The *in vitro* growth kinetics of SBV were determined in KC and Aag2 cells following infection with an MOI of 10 for 1 h. Samples were collected at 0, 8, 24, and 48 h postinfection, and virus titer was determined by plaque assays in CPT-Tert cells. Each experiment was performed in triplicate and repeated twice.

Western blotting. Protein expression in BTV-1-infected KC cells was assessed by SDS-PAGE and Western blotting using polyclonal rabbit antiserum raised against NS1 (41) and a rabbit polyclonal anti-actin antibody as the control (Sigma; A5060).

Detection of proteins by immunofluorescence. Formaldehyde-fixed cells were permeabilized by incubation in 0.3% Triton-PBS for 30 min, followed by washing with PBS, a further incubation in 0.1% SDS-PBS for 10 min, and incubation in PBS. Cells were preincubated in CAS-Block for 1 h at room temperature, followed by an incubation with CAS-Block-diluted polyclonal rabbit antiserum raised against BTV-1 NS1 (1:2,000) (41) or SBV N (50) (1:500) for 90 min at 37°C. Cells were then washed three times for 5 min with PBS. Following this, an anti-rabbit antibody conjugated with Alexa 488 diluted in CAS-Block (1:3,000) was added and incubated for 1 h at 37°C. Following a further washing step with PBS, cells were dried and covered with 4′,6-diamidino-2-phenylindole (DAPI) containing mounting medium (Vectashield; hard set), and fluorescence was detected using a Zeiss LSM meta microscope.

Small RNA isolation and deep sequencing analysis. Sequencing of small RNAs was performed by using the Illumina Solexa platform; KC (8×10^6/well) and Aag2 (2.6×10^6/well) cells in 6-well plates were infected by BTV-1 at an MOI of 0.2 or SBV at an MOI of 10. At 24 h postinfection (KC cells) or 48 h postinfection (Aag2 cells), total RNA was isolated using 1 ml TRIzol (Invitrogen) per well. Glycogen was added prior to isopropanol addition to enhance recovery of small RNA from samples. Total RNA was then loaded on a 15% denaturing urea acrylamide gel, and RNA molecules of 18 to 30 nt in size were purified from the gel, linked to adapters, reverse transcribed, and sequenced by ARK-Genomics (The Roslin Institute, University of Edinburgh) on the Illumina Solexa platform (HiSeq 2000). Illumina adapters and sequencing primers were removed using cutadapt (51), and the trimmed sequences were aligned to the reference genome using Novoalign. Graphs and reports were produced in R (52) using the viRome package (http://www.ark-genomics.org/services-bioinformatics/virome). The complement-distance plots were calculated as follows: the distance between the 5′ end of reads of 24 to 30 bp that map on complementary strands was counted, and the sum of counts was plotted against the distance. For the sequence logos, counts of each base at each position were used to create a position-weight matrix, and the subsequent sequence logo was plotted using the seqLogo (53) package from Bioconductor (54). Small RNAs were mapped to SBV (GenBank accession numbers JX853179 to JX853181) or BTV-1 (GenBank accession numbers JX680457 to JX680466).

Nucleotide sequence accession number. Small RNA sequences (of data shown and repeats) were submitted to the European Nucleotide Archive (accession number ERP001936).

RESULTS

An active RNAi pathway is present in the KC midge cell line. An RNAi response can be induced in arthropod cells by sequence-specific dsRNA either by transfection, or in case of drosophila cells, by addition to the cell culture medium (55). We designed a luciferase-based reporter assay in order to investigate if *C. sonorensis*-derived KC cells can induce a dsRNA-mediated RNAi response. As little is known about the efficiency and/or toxicity of transfection reagents in KC cells, a pilot experiment was carried out with either fluorescently labeled dsRNA molecules or plasmid DNA and by using different transfection reagents. At 24 h post-seeding, KC cells were incubated with fluorescein-labeled dsRNA or fluorescein-labeled plasmid DNA in the absence or presence of different transfection reagents (Fugene HD, GeneJammer, or Lipofectamine 2000) by following the manufacturer's protocol. At 24 h postincubation, we estimated the number of transfected cells (green cells) by fluorescence microscopy, with at least 400 cells counted for each condition (data not shown). No obvious toxicity was detected 24 h posttransfection (data not shown). Use of Fugene and GeneJammer resulted in similar numbers of transfected cells (between 14% and 23%). Lipofectamine 2000-mediated transfection resulted in the highest number of transfected cells for dsRNA (approximately 40%), similarly to what we obtained in cells incubated with dsRNA in the absence of transfection reagent (32.5%). In contrast, transfection of KC cells using Lipofectamine 2000 gave <5% of transfected cells after plasmid DNA transfection, compared to 27% in GeneJammer and 11% in Fugene. No

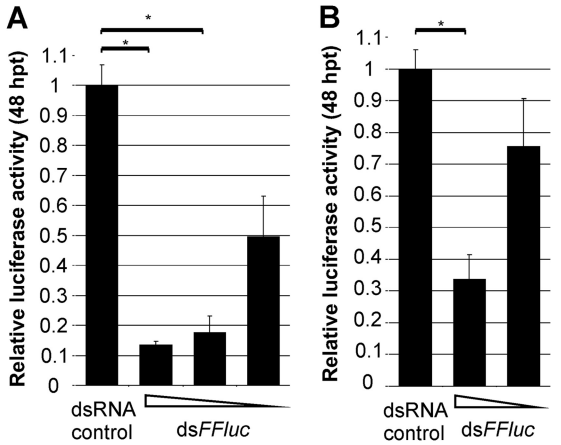


FIG 1 KC cells have a functional dsRNA-inducible RNAi pathway. KC cells were cotransfected with plasmids pAcIE1-Rluc and pIZ-Fluc. Following transfection, different concentrations of either eGFP-specific (ds-control) or *FFluc*-specific (ds*FFLuc*) dsRNA were either transfected (A) or added to the culture media 24 h posttransfection of plasmids (B). The relative luciferase expression (*FFluc/RLuc*) was determined at 48 h posttransfection (hpt). The mean with standard error is shown for two independent experiments performed in triplicate. *, $P < 0.05$, *t* test.

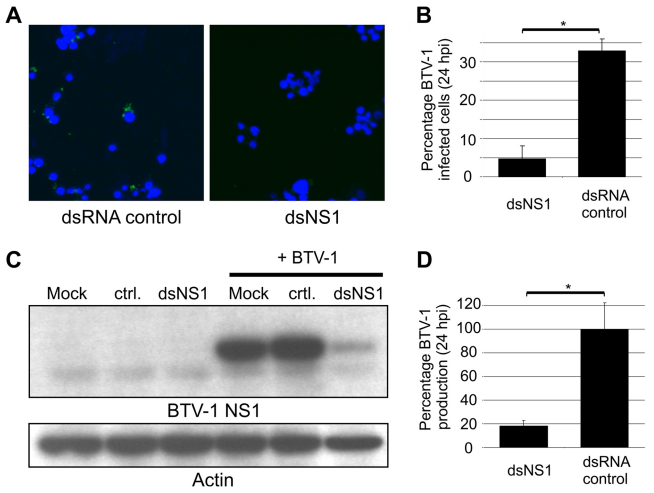


FIG 2 BTV-1 can be targeted by dsRNA in KC cells. (A) KC cells were transfected with either eGFP-specific control dsRNA or dsRNA targeting BTV NS1 (dsNS1), followed by BTV-1 infection at 24 h posttransfection. Twenty-four hours postinfection (hpi), cells were fixed and BTV-infected cells were visualized using a BTV NS1-specific primary antibody and an anti-rabbit secondary antibody conjugated to Alexa 488 (green signal). Cell nuclei were stained with DAPI. (B) Quantification of fluorescent cells. Infected and noninfected cells were counted, and the respective percentages from three independent experiments for each treatment were determined. (C) BTV-1 NS1 expression was determined by Western blot analysis in lysates of dsRNA-transfected and BTV-1-infected KC cells at 24 h postinfection using a BTV NS1-specific primary antibody. Western blot detection of actin was used as a loading control. (D) BTV-1 titers in cell culture supernatant, determined 24 h postinfection. Means from three independent experiments are shown; error bars represent standard errors. *, $P < 0.05$, *t* test.

green fluorescing cells were observed if plasmid DNA was added to cells in the absence of a transfection reagent.

Next, we determined the ability of dsRNA, either added to the medium or transfected with GeneJammer, to silence a firefly luciferase (*FFluc*) reporter gene. Cells were cotransfected with plasmids expressing *FFluc* as well as *Renilla* luciferase (*RLuc*) as an internal control expressed from baculovirus promoters (OpIE2 for *FFluc* in pIZ-Fluc and AcIE1 for *RLuc* in pAcIE1-Rluc). At 24 h postseeding, we induced an RNAi response by transfection (Fig. 1A) or addition (Fig. 1B) of different concentrations of either eGFP-specific control dsRNA (ds-control) or dsRNA targeting *FFluc*. We assessed luciferase activities 24 h after the addition of dsRNA. A concentration-dependent reduction of luciferase activity was found for cells treated with *FFluc*-specific dsRNA, regardless of whether dsRNA was transfected (Fig. 1A) or added to the cell culture medium (Fig. 1B). These results show that KC cells are able to induce a sequence-specific, dsRNA-dependent RNA silencing response. In addition, our data show that, similarly to what has been reported for Schneider-2 (S2) drosophila cells (56, 57), KC cells are able to take up dsRNA from culture medium.

The dsRNA-inducible RNAi response in midge cells displays antiviral activity. We next investigated the ability of this pathway to inhibit virus replication. As BTV is known to be transmitted to susceptible mammals via *Culicoides* species (27), we used this virus to investigate the dsRNA-induced antiviral RNAi response in KC cells. We first transfected the KC cells with dsRNA targeting BTV-1 NS1 or eGFP-specific control dsRNA. After 24 h, we infected the cells with BTV-1 at a multiplicity of infection (MOI) of 0.2. The success of silencing was assessed at 24 h postinfection by immunofluorescence (Fig. 2A) and Western blot detection (Fig. 2C) using an NS1-specific antibody. A reduction in NS1-positive cells was detected when dsRNA targeting NS1 was transfected and compared to control dsRNA. Approximately 35% of control dsRNA-transfected cells expressed BTV-1 NS1 as assessed by fluorescence microscopy, in contrast to 15% when dsRNA targeting NS1 was transfected (Fig. 2B). We confirmed these results by Western blotting, where levels of NS1 expression were greatly reduced in cells transfected with virus-specific dsRNA (Fig. 2C). BTV NS1 has been shown to be important for viral replication (58, 59), and consequently knockdown of NS1 expression will have a negative effect on virus production. We therefore measured BTV infectious viral particles released in the supernatant of KC cells incubated with dsRNA (targeting BTV-1 NS1 or control) at 24 h postinfection. As expected, a significant decrease in BTV-1 production was detected in cells incubated with dsRNA targeting NS1 compared to that of control infections (Fig. 2D). Similar results were obtained when we added (rather than transfected) dsRNA to the culture medium (data not shown), suggesting a capability for antivirally active dsRNA uptake similar to that observed for *Drosophila* cell lines (56, 57).

***Culicoides* cells can mount an RNAi response against the dsRNA bluetongue virus.** Having shown that RNAi can be induced in KC cells following dsRNA transfection, we investigated whether such a response could be induced following viral infection. Induction of an antiviral RNAi response is characterized by the production of small RNA molecules that map to the viral genome and/or antigenome (5, 6, 11, 12, 19, 20, 60). We isolated total RNA from KC cells 24 h postinfection with BTV-1, and we then sequenced small RNAs below 40 nt by Illumina Solexa sequencing and determined the sequences, frequencies, and BTV genome location. For most BTV-1 segments, the viRNA molecules produced in infected KC cells were predominantly 21 nt in length and mapped to both the coding and noncoding strand with similar frequencies (Fig. 3

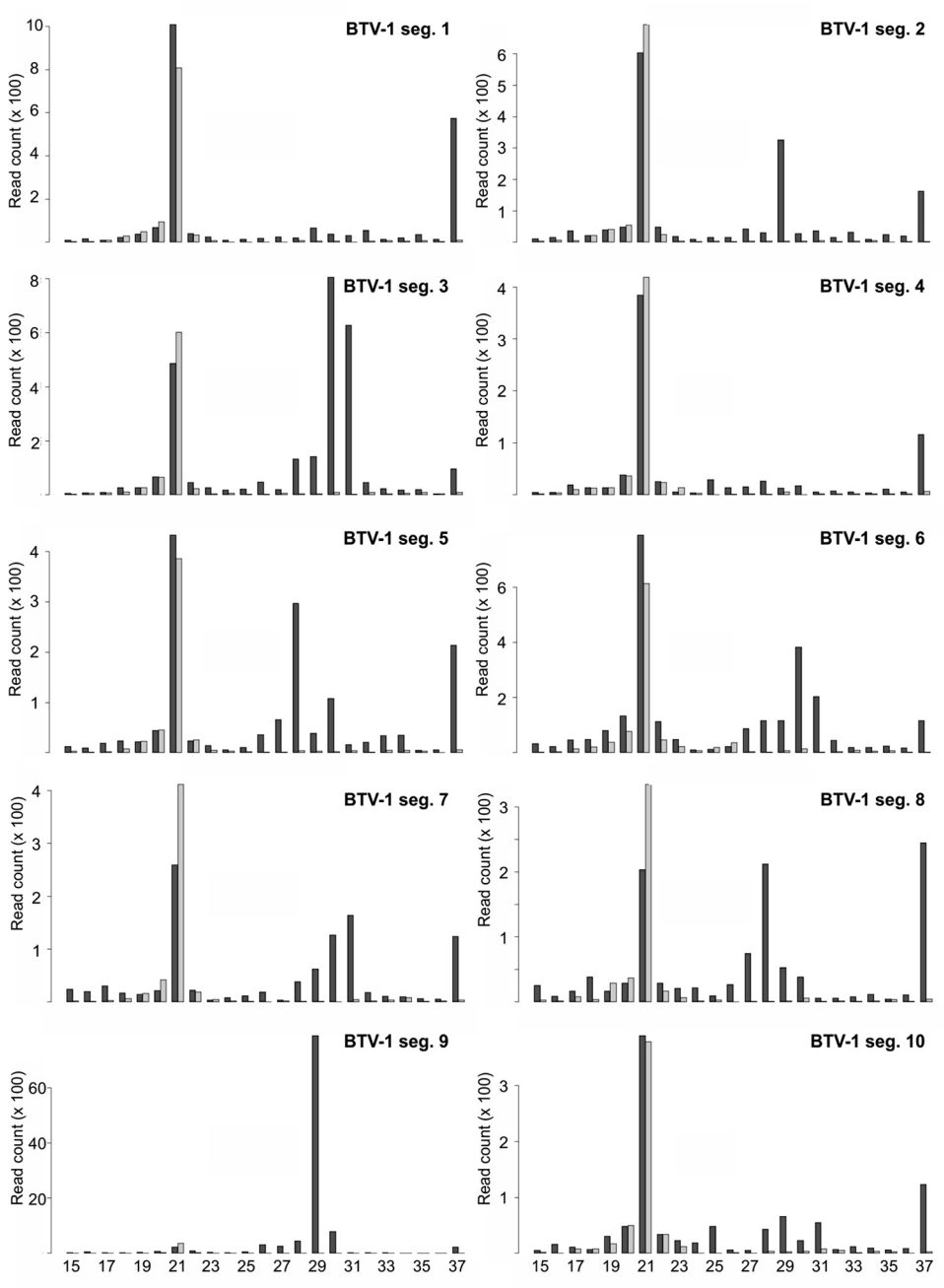


FIG 3 BTV-1 is targeted by the RNAi response in KC cells. Size distribution of small RNA molecules mapping to each segment (1 to 10) of BTV-1 in KC cells at 24 h postinfection. The *y* axis indicates the frequency of small RNAs; the *x* axis indicates the length in nucleotides. Dark grey indicates small RNAs mapping to the coding strand, and light grey indicates small RNAs mapping to the noncoding strand.

and 4). The 21-nt viRNAs were distributed along the coding and noncoding genome strand with variable frequency and a hot (high viRNA reads) and cold (low or no viRNA reads) spot distribution, similar to what has been described for mosquito-borne arboviruses (5, 6, 11, 12, 19, 20, 60). The frequency of the viRNAs was segment dependent. In addition to 21-nt viRNAs, other classes of small RNAs between 26 to 31 nt in length with a bias for the BTV-1 coding strand were also identified. The frequency of these longer RNAs differed per segment from few (segment 1) to the majority of virus-specific small RNAs (segment 9) (Fig. 3 and 4). To determine if these results were specifically induced by BTV, we investigated the produced small RNAs against BTV-1 in the nonvector cell line Aag2 (derived from *Aedes aegypti*). We established by fluorescence microscopy that BTV-1 was able to infect Aag2 cells at levels comparable to KC cells (Fig. 5). RNA was isolated 48 h postinfection, and small RNAs were sequenced on the Illumina sequencing platform as described before. The viRNA production pattern,

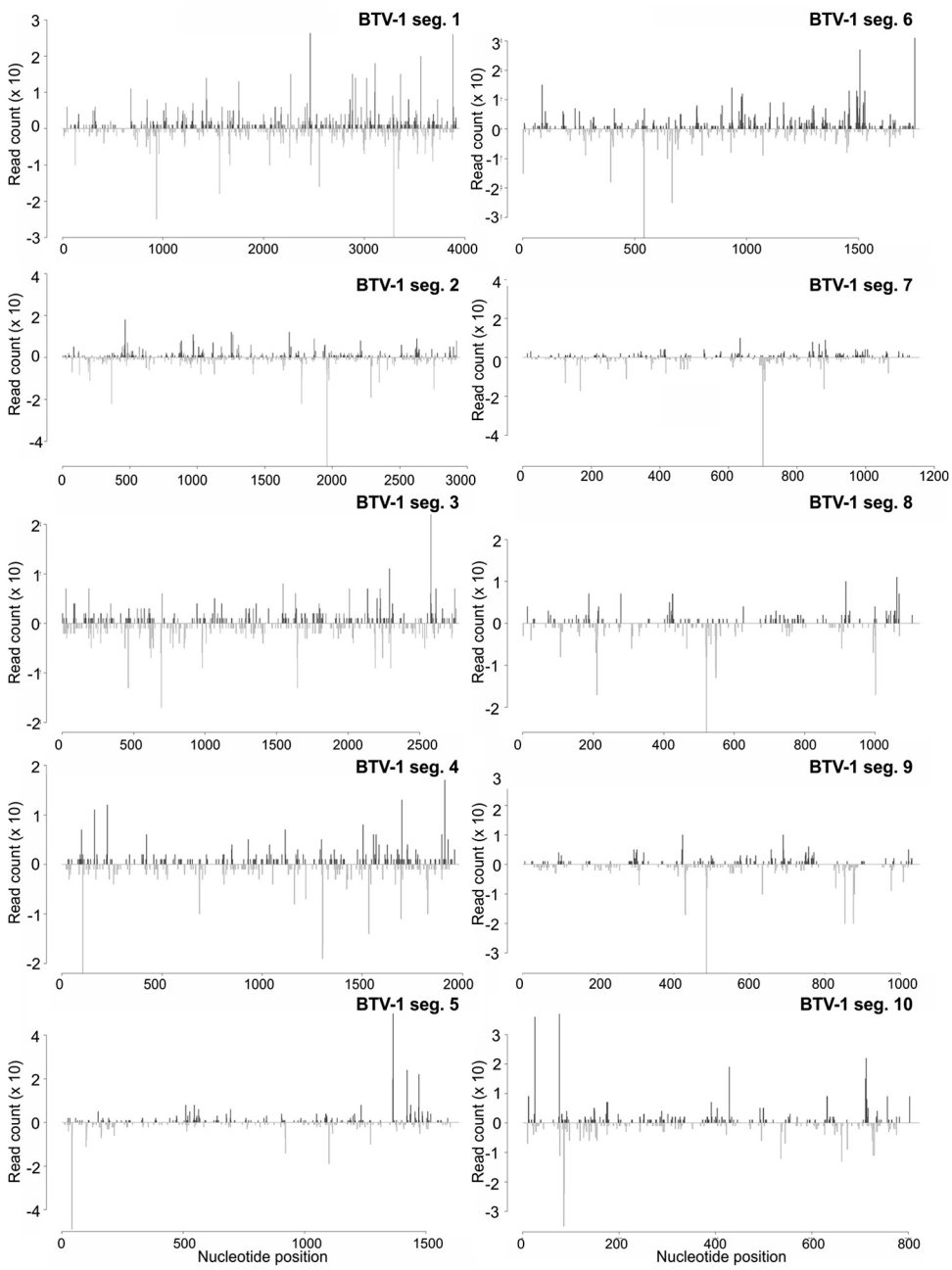


FIG 4 Distribution of 21-nt viRNAs of BTV-1 in KC cells. Frequency distribution of 21-nt viRNAs to each segment (1 to 10) of BTV-1. The *y* axis shows the frequency of the 21-nt viRNAs mapping to the corresponding nucleotide position on the *x* axis. Positive numbers and peaks represent the frequency of viRNAs mapping to the coding strand (in 5′→3′ orientation). Negative numbers and peaks represent those viRNAs mapping to the noncoding strand (in 3′→5′ orientation).

including the larger class of virus-specific small RNAs, was largely conserved in Aag2 cells infected with BTV-1 (Fig. 6 and 7). Importantly, pattern, location, and frequencies of viRNA production were conserved mainly between independent experiments (data not shown). Together, these data show that the antiviral RNAi responses following infection by BTV of *C. sonorensis*-derived KC cells and *Aedes aegypti*-derived Aag2 cells are broadly comparable and are predominantly characterized by the production of 21-nt viRNAs for most of the segments.

***Culicoides* cells can mount an RNAi response against the negative-strand RNA Schmallenberg virus.** SBV is a recently emerged pathogen belonging to the *Orthobunyavirus* genus of the *Bunyaviridae* family and is thought to be most probably midge-borne (4). We infected KC cells with SBV, and RNA was isolated 24 h postinfection to determine if a similar production pattern of viRNA production is observed in KC cells infected with an arbovirus belonging to a different virus family from the *Reoviridae*. Infection experiments indicated that SBV infects and replicates

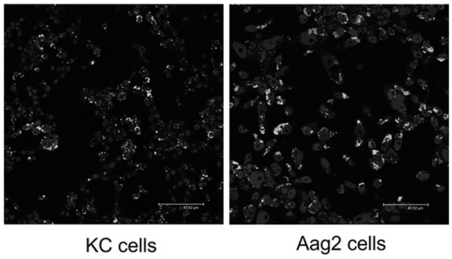


FIG 5 BTV-1 infects nonvector mosquito Aag2 cells. BTV-infected KC and Aag2 cells were fixed at 48 h postinfection, and expression of NS1 protein was monitored by immunofluorescence (bright signal). Cell nuclei were stained with DAPI.

with low frequency in KC cells even after high multiplicity of infection (Fig. 8 A and B). Similar experiments performed on Aag2 cells showed enhanced infection and replication of SBV (Fig. 8A and B). Therefore, SBV-infected Aag2 cells were used in order to investigate if the results obtained for BTV-1-infected KC and Aag2 cells could be broadened to another arbovirus. Sequences, frequencies, and SBV genome locations of small RNAs below 40 nt were determined as described above. As shown in Fig. 8C, 21-nt viRNAs are the predominant species of viRNA in KC cells and for most of the segments in Aag2 cells. Again, a distribution along the L, M, or S genome and antigenome with hot and cold spots was observed (Fig. 9A). In both KC and Aag2 cells, most viRNA reads are generated by the S segment, followed by the M and L segments (Fig. 8C). In addition to 21-nt viRNAs, small RNAs in the size range of 24 to 30 nt with a bias for the positive antigenome strand were detected in Aag2-infected cells and matched with all three viral segments, although at different frequencies. In the case of the S segment, these larger small RNAs represented the majority of small RNAs (Fig. 8C). As described above, similar class sizes of small RNAs were also detected in BTV-1-infected KC and Aag2 cells. This size range of small RNAs normally represents the group of PIWI-interacting RNA (piRNA) molecules, known to be important in suppressing transposons in germ line cells in various organisms, including drosophila, zebrafish, and mice (61–65). Primary piRNAs are normally antisense to the genomic regions (mostly transposons) and target transposon-derived single-stranded sense RNA. Upon cleavage, secondary piRNAs are produced that are mostly sense and used to find complementary antisense RNA, resulting again in primary-type piRNAs. Recently, virus-specific piRNA-like molecules have been reported for several arboviruses, including CHIKV, Sindbis virus (SINV), and LACV in aedine mosquitoes or their derived cell lines (6, 18, 19, 44). Due to the so-called "ping-pong" mechanism of piRNA production, piRNAs have specific features. The primary piRNAs are in antisense orientation and have a bias for uridine at position 1. In contrast, secondary piRNAs are in the sense orientation and have an adenine at position 10. In addition, complementary piRNAs and viral piRNA-like molecules of LACV and SINV are often separated at the 5′ end by 10 nucleotides (44, 66). Most of the SBV-specific small RNAs of 24 to 30 nt produced in Aag2 cells have piRNA-specific features (sense [antigenome] with A_{10} and antisense [genome] with U_1 [Fig. 10A] and separation of complementary RNAs at the 5′ end by 10 nucleotides [Fig. 10B]), in particular those produced from the M and S segments. They are distributed along the segments and do not map to a specific region of the segments (Fig. 9B). In contrast, the BTV-1-specific 24- to 31-nt small RNAs produced in KC and Aag2 cells do not show any specific sequence logo (data not shown). Pattern, location, and frequencies of viRNA production were conserved mainly between independent repetitions (data not shown).

Taken together, our data show that midge-derived KC cells mount an antiviral RNAi response following infection with arboviruses with different genome structures. Regardless of the virus infecting these cells, viRNAs of 21 nt in length were found to be the dominant class of virus-specific small RNA. Thus, the *Culicoides* antiviral RNAi response resembles mainly similar pathways found in mosquitoes. Larger small RNAs with features of piRNA-like molecules were found for SBV-infected Aag2 cells but not KC cells. Although similar RNA molecules were detected in BTV-infected KC and Aag2 cells, these RNAs do not possess piRNA-specific features.

DISCUSSION

RNAi (and in particular the exogenous RNAi pathway) has been shown to be a major antiviral response against arboviruses in mosquitoes and an important process regulating this virus/host interaction (13, 14). *Culicoides* midges are one of the major invertebrate vectors of several arboviruses of humans and livestock. In this study, we have shown that *C. sonorensis*-derived KC cells mount an antiviral RNAi response.

Drosophila cells have been reported to take up dsRNA from culture medium, a phenomenon not observed for any of the mosquito-derived cell lines (56, 57). Interestingly, we show that also KC cells are able to take up dsRNA molecules directly from the culture medium, although the precise pathway for dsRNA uptake is not known. Several genes have been linked with the dsRNA uptake in *D. melanogaster*, suggesting receptor-based endocytosis (56, 57, 67). Due to the lack of genomic information on *Culicoides*, it is currently difficult to draw any comparisons.

dsRNA and siRNA molecules have been used to target and silence a variety of viruses in numerous organisms. Even mammalian cells, believed to not naturally mount a siRNA-based antiviral RNAi response upon viral infection, can induce an siRNA-based antiviral RNA silencing response after transfection of siRNA molecules (68). Therefore, successful inhibition of virus production following transfection of dsRNA targeting viral sequences cannot be used solely as an indication that such an antiviral RNA silencing response occurs during natural infection. A key feature of antiviral RNAi in mosquitoes (as in other insects) is the production of 21-nt viRNA molecules (13, 14). Deep sequencing of KC cells infected with two different midge-borne arboviruses (BTV and SBV) shows that the majority of viRNAs for most BTV genomic segments (8 out of 10) and all SBV segments are 21 nt in length. viRNAs in mosquito cells (and other insects) are produced by Dcr-2 cleavage of dsRNA molecules that could derive either from replicative intermediates or secondary structures within the viral transcripts (14) but also the viral genome itself in case of dsRNA viruses such as BTV. From the available viRNA profiles, replication intermediates are generally the favored Dcr-2 substrate candidates. This is supported by the scattering of viRNAs across the whole genome/antigenome, the presence of hot and cold spots of viRNA production (5, 6, 11, 12, 19, 20, 69), and no real preference of a longer stretch of only the genome or antigenome as the producer of 21-nt RNAs, which would be expected if secondary RNA structures are the favored Dcr-2 substrate. Recently, it has been

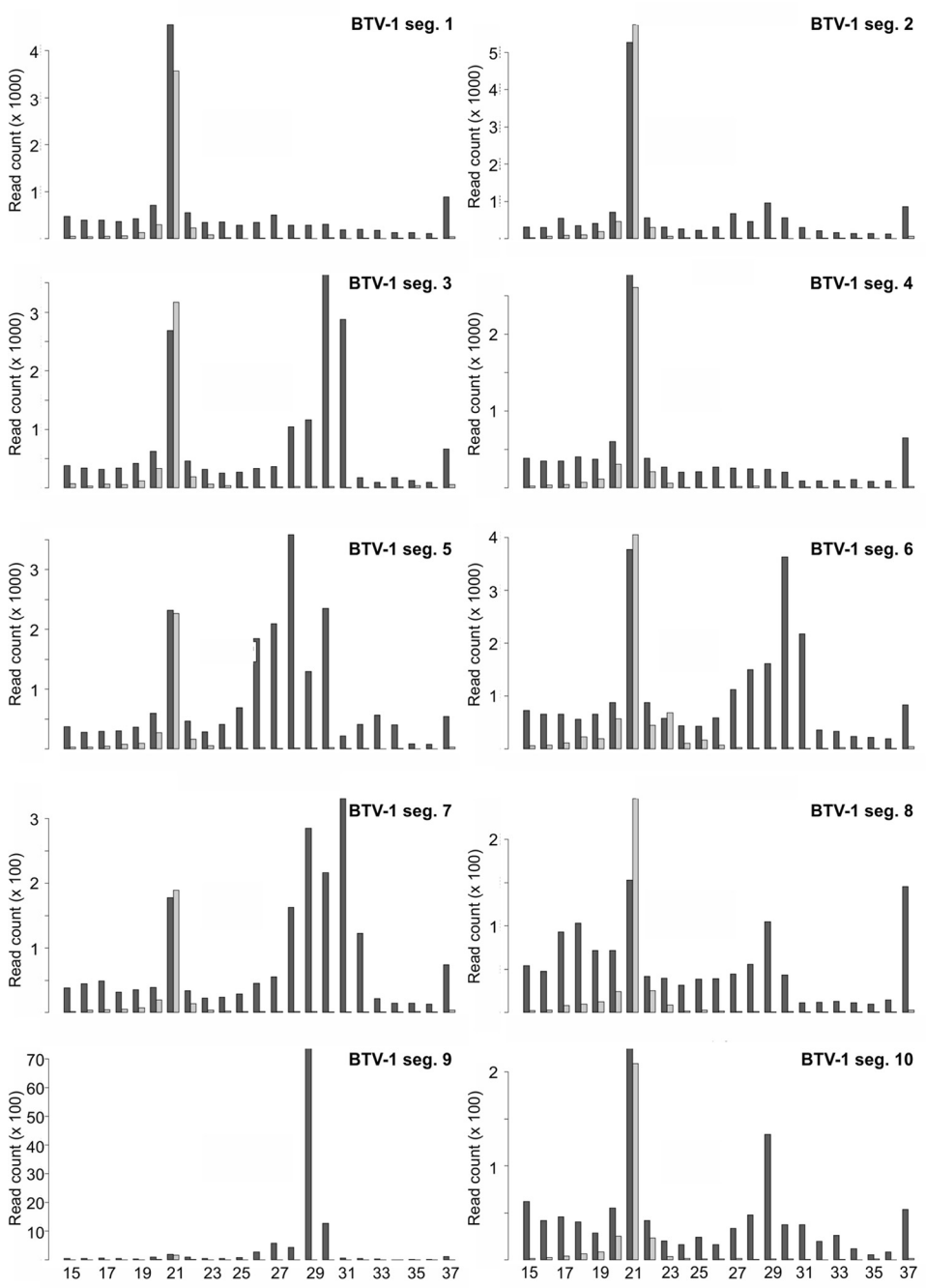


FIG 6 BTV-1 infection is targeted by the RNAi response in Aag2 cells. Size distribution of small RNA molecules mapping to each segment (1 to 10) of BTV-1 in Aag2 cells at 48 h postinfection. The *y* axis indicates the frequency of small RNAs; the *x* axis indicates the length in nucleotides. Dark grey indicates small RNAs mapping to the coding strand, and light grey indicates small RNAs mapping to the noncoding strand.

shown that certain RNAs can be over- or underrepresented in small RNA libraries, due to low sequencing depth and cloning bias (70, 71). Some of the observed hot and cold spots could be the result of such a cloning bias; however, the presence of small RNAs mapping to the noncoding strand of BTV with a similar frequency as to the coding strand strongly supports the dsRNA genome as the RNAi inducer molecule. Before this study, it was not immediately apparent whether an RNAi response of 21-nt viRNAs would be induced following infection by viruses with a dsRNA genome like the orbivirus BTV and if the inducer molecules would be the dsRNA genome or secondary structures in the viral transcripts. Considering that synthesis of the negative-sense RNA during the viral replication cycle is believed to occur only within the newly assembled viral particles, the secondary structures of the viral transcripts would be the favored RNAi inducer molecule; this strategy helps dsRNA viruses also to shield them from host pro-

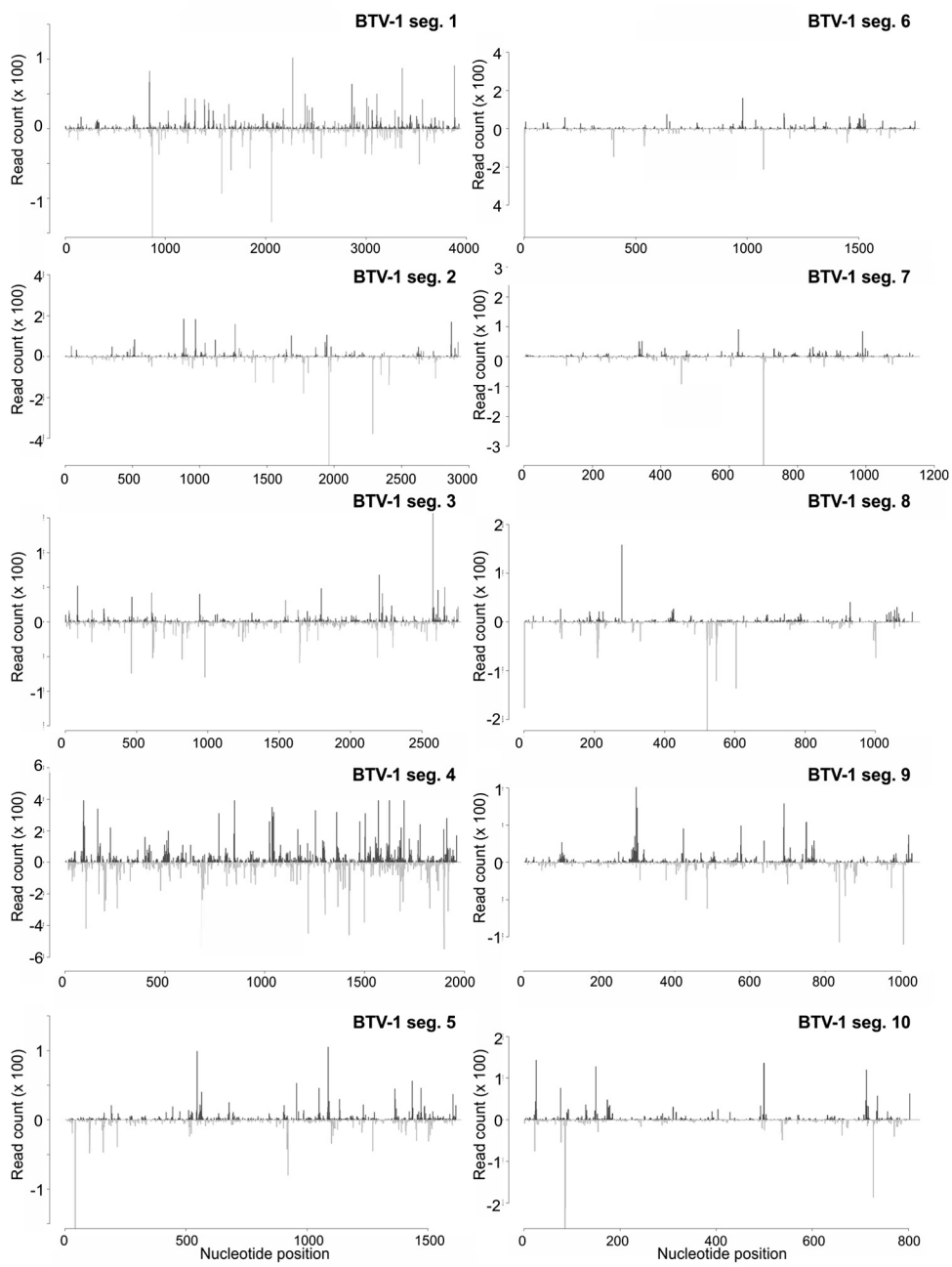


FIG 7 Distribution of 21-nt viRNAs of BTV-1 in Aag2 cells. Frequency distribution of 21-nt viRNAs in infected Aag2 cells to each segment (1 to 10) of BTV-1. The *y* axis shows the frequency of the 21-nt viRNAs mapping to the corresponding nucleotide position on the *x* axis. Positive numbers and peaks represent the frequency of viRNAs mapping to the coding strand (in 5′→3′ orientation). Negative numbers and peaks represent those viRNAs mapping to the noncoding strand (in 3′→5′ orientation).

teins such as RNA sensors that activate the antiviral interferon pathway in mammals (72–76). Thus, viRNAs of dsRNA viruses could have been solely derived from the secondary structures of the coding RNA strand (which functions as mRNA), or these viruses could escape antiviral RNAi altogether (72, 75, 76). The features of BTV viRNAs detected in our study (scattering across the whole genome in sense and antisense patterns) indicate that at least a small amount of dsRNA viral genome is accessible to the RNAi machinery and probably not protected by the double-layered viral membrane particles, as recently shown for the induction of interferon (IFN) in mammalian cells by BTV dsRNA (77). These results are in line with the detection of viRNAs derived from dsRNA viruses (the birnavirus drosophila X virus [DXV] as well as drosophila totivirus) (78) during persistent infection in a drosophila cell line and the increase in susceptibility for DXV in RNAi-deficient drosophila (79). The resulting viRNAs are expected to be able to target the BTV transcripts present in the cytoplasm, resulting in less viral protein production and subse-

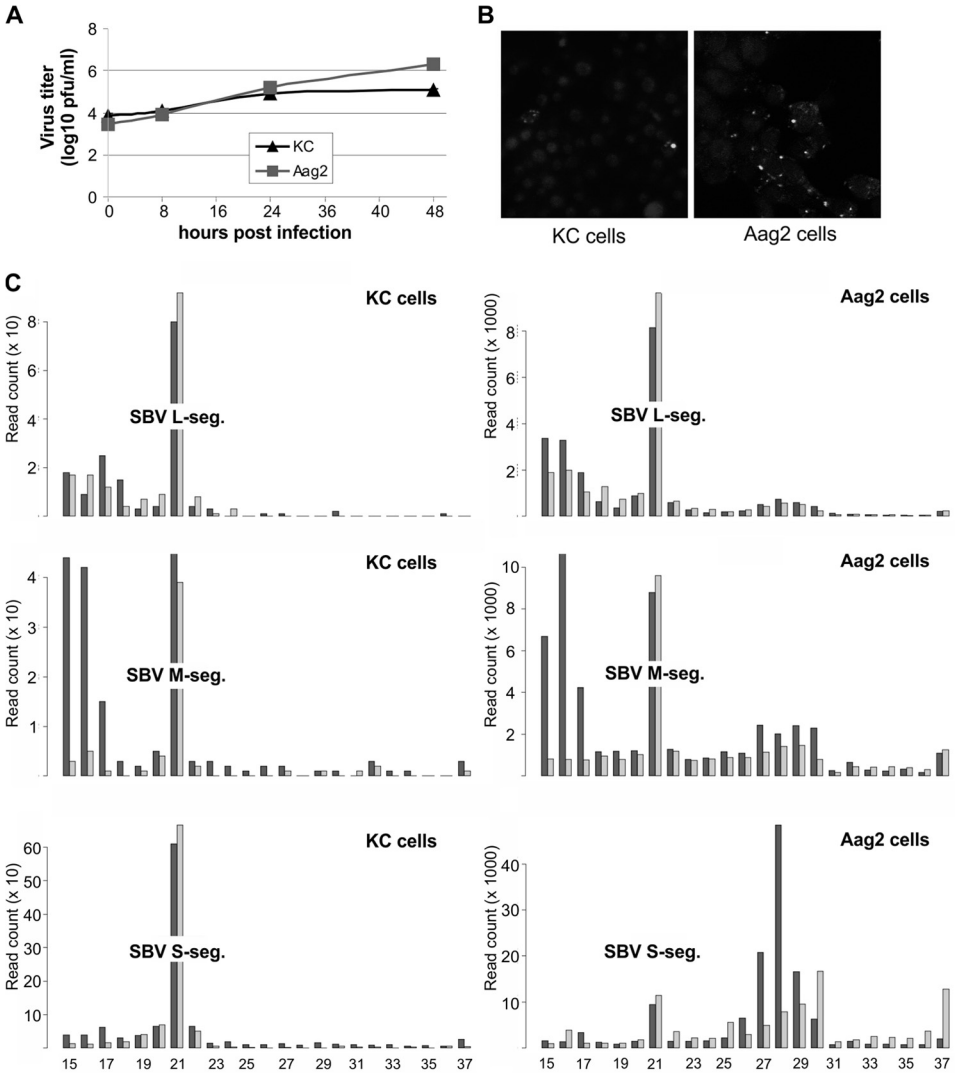


FIG 8 SBV infection and targeting by the RNAi response in KC and Aag2 cells. (A) Virus production in supernatant of KC (triangle) and Aag2 (square) cells infected with SBV at a MOI of 10 was determined at various time points postinfection (0, 8, 24, and 48 h) by plaque assay. A representative of two independent experiments performed in triplicate is shown; standard errors are indicated. (B) SBV-infected KC and Aag2 cells were fixed at 48 h postinfection, and SBV-infected cells were visualized by using an SBV N-specific primary antibody (bright signal). Cell nuclei were stained with DAPI. (C) Size distribution of small RNA molecules mapping to L, M, and S of SBV in KC cells at 24 h postinfection or in Aag2 cells at 48 h postinfection. The *y* axis indicates the frequency of small RNAs; the *x* axis indicates the length in nucleotides. Dark grey indicates small RNAs mapping to the coding strand, and light grey indicates small RNAs mapping to the noncoding strand.

quently reduced virus titers, as shown in this study. In addition to the 21-nt viRNAs, larger classes of BTV-specific small RNAs of 26 to 31 nt in length were detected. This resembles the size distribution of piRNAs, a class of Dicer-independent small RNAs found in vertebrate and invertebrates thought mainly to be important for genome stability in germ line cells by targeting transposons (61–65). Recently, virus-specific piRNA-like small RNAs were found in arbovirus-infected aedine mosquitoes and derived cell lines. These piRNA-like molecules were found to map mainly to the coding strand of the viral genome of positive-strand RNA arboviruses but also to the antigenome of LACV (18, 19, 44). Due to their production pathway (the so-called ping-pong amplification mechanism), piRNAs and viral piRNA-like molecules have a specific sequence logo (61–65). BTV-specific small RNAs of 26 to 31 nt in length do not show a piRNA-like sequence logo (not shown). It is not known how these larger BTV-specific RNA molecules are produced, or their function, or if they are related to RNAi processes at all. Their production is virus specific and not cell type specific, as we also detected them in the Aag2 mosquito cell line.

In the case of SBV, 21-nt viRNAs could be detected in both KC and Aag2 cells for all three segments in an asymmetric distribution, suggesting again a dsRNA replication intermediate as the inducer of the RNAi response. This is in line with results observed in drosophila cells infected with LACV (6). Compared to infection with positive-strand RNA viruses or dsRNA viruses, no dsRNA could be detected for negative-strand RNA viruses, at least in infected vertebrate cells (80). Our results indicate that dsRNA replicative intermediates are present in orthobunyavirus-infected cells, possibly at low levels but levels still sufficient to induce an antiviral RNAi response. Longer small RNA molecules containing

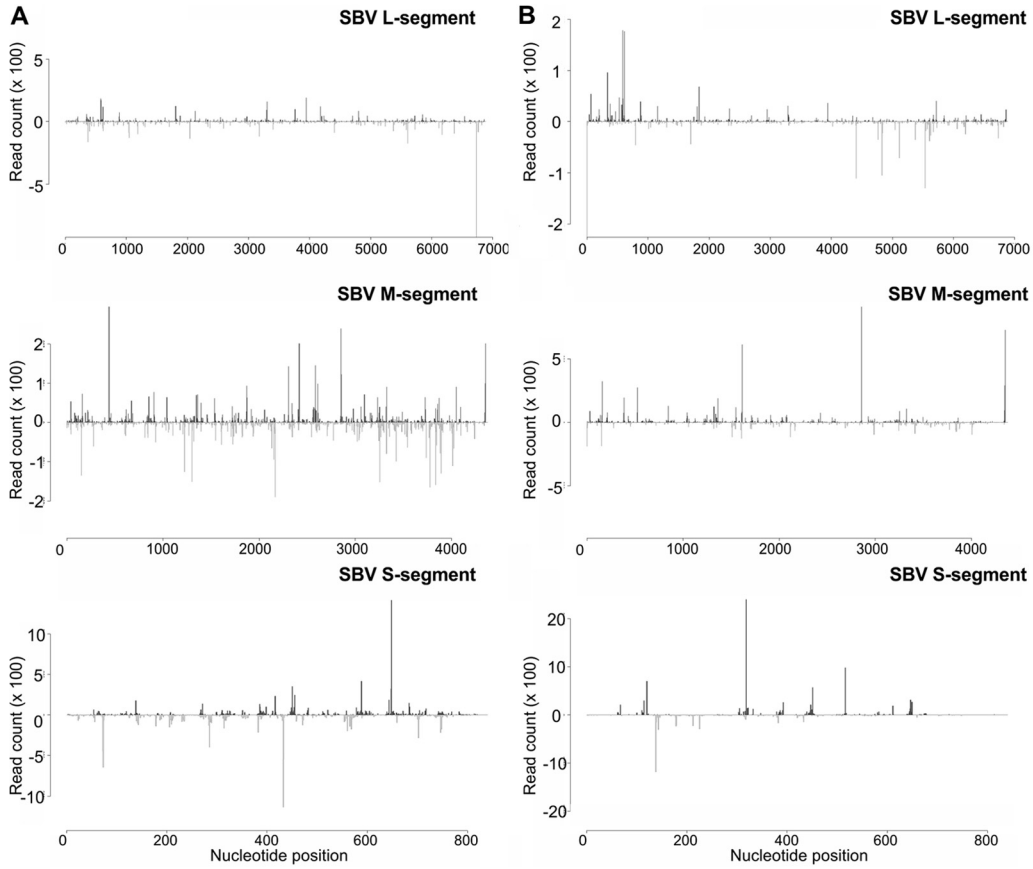


FIG 9 Distribution of SBV-derived small RNAs in Aag2 cells. Frequency distribution of 21-nt viRNAs (A) or 24- to 29-nt small RNAs (B) to L, M, and S of SBV in Aag2 cells at 48 h postinfection. The *y* axis shows the frequency of the 21-nt viRNAs mapping to the corresponding nucleotide position on the *x* axis. Positive numbers and peaks represent the frequency of viRNAs mapping to the coding strand (in 5′→3′ orientation), and negative numbers/peaks indicate those viRNAs mapping to the noncoding strand (in 3′→5′ orientation).

piRNA features were also detected in SBV-infected Aag2 cells, as recently described for other bunyaviruses (LACV, Rift Valley fever virus [RVFV]) in RNAi-deficient *Aedes albopictus*-derived C6/36 cells and other arboviruses (such as CHIKV, SINV, DENV, and RVFV) in aedine mosquitoes and/or their derived cell lines (18, 19, 44, 81). As no piRNA-like molecules could be detected in any of the BTV-1-infected cell lines, production of piRNA-like molecules may be specific to single-stranded RNA viruses, though more research is needed to answer this question. It is not known if these piRNA-like molecules have an antiviral function and how they are induced. Neither BTV nor SBV produced piRNA-like molecules in KC cells. This could be due to either the lack of a piRNA pathway in *Culicoides* species or, alternatively, a deficiency of the KC cell line.

The detection of BTV- or SBV-specific 21-nt viRNAs indicates the ability of KC cells to induce an antiviral RNAi response. In the absence of any genomic information on *Culicoides*, we can only speculate that orthologs of exogenous RNAi pathway proteins such as Dcr-2 or Ago-2 are present in the midge genome. However, the presence of 21-nt viRNAs is a strong indicator for the presence of an exogenous RNAi pathway that is comparable to that of mosquitoes (14). This raises a number of questions: most importantly, how is BTV or SBV able to successfully replicate and infect its midge vectors? Plant and true insect viruses have been reported to encode proteins able to interfere with the antiviral RNA silencing response by expressing RNA silencing suppressor proteins (RSS) (23, 82). Until now, no RSS protein has been identified in arboviruses, suggesting other strategies for inhibition or evasion of the RNAi response. Previous data obtained from SFV infection of mosquito cells suggests a decoy mechanism: viRNAs that are produced at high concentrations are not able to target the virus efficiently in contrast to viRNAs produced at low concentrations that result in efficient silencing of the virus. This strategy results in successful replication at least for some time postinfection, even though viRNAs are produced (12). The mosquito-borne flaviviruses WNV and DENV have been recently shown to express a subgenomic flavivirus RNA able to interfere with the RNAi response in mosquito cells, thereby ensuring efficient viral replication (83). However, all mosquito-borne arboviruses investigated thus far are efficiently targeted by the RNA silencing response (5, 6, 11, 12, 19, 20, 60). Further research is needed to determine if and whether SBV or BTV do have any (even if weak) RNAi evasion/inhibition strategies.

Taken together, our findings define for the first time the presence of an RNAi response in *Culicoides* cells which is able to target midge-borne arboviruses and resembles at least in part the exogenous antiviral RNAi pathway of mosquitoes. More work will be

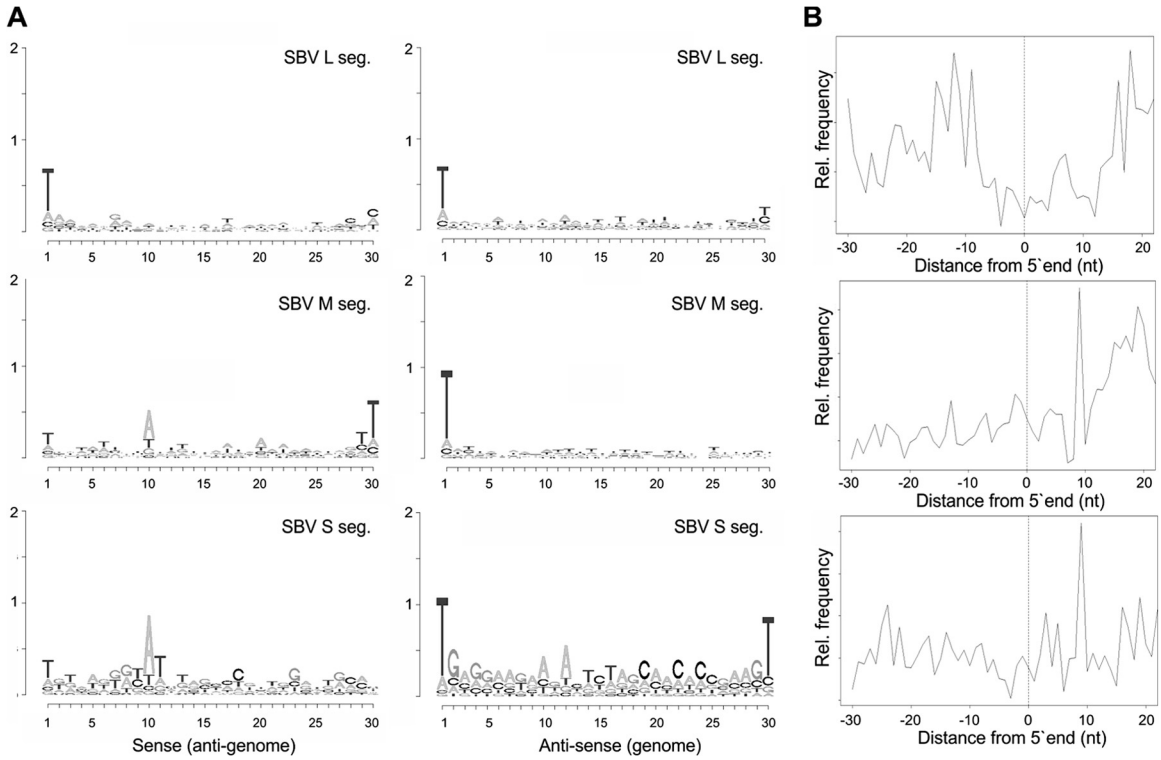


FIG 10 SBV-specific piRNA-like molecules produced by viral infection in Aag2 cells. (A) Relative nucleotide frequencies and conservations per position of 24- to 30-nt small RNAs mapping to the antigenome (left) and genome (right) of SBV in Aag2 cells are indicated for L, M, and S. The overall height of the nucleotide represents the sequence conservation. Note: the sequence is represented as DNA. (B) Frequency mapping of the distance of complementary 24- to 30-nt small RNAs mapping to SBV segments (top to bottom: L, M, and S) in Aag2 cells.

required to determine the exact mechanisms and proteins involved in the RNAi pathway in midges, but this study allows further investigations into these processes.

ACKNOWLEDGMENTS

This work was supported by United Kingdom Medical Research Council (A.K.), Wellcome Trust (M.P., M.V., R.M.E.), and Netherlands Organization for Scientific Research NWO (Rubicon Fellowship) (E.S.).

We thank M. Beer (Friedrich-Loeffler-Institut, Germany) for providing the SBV sample.

REFERENCES

1. **Weaver SC, Reisen WK.** 2010. Present and future arboviral threats. Antiviral Res. **85**:328–345.
2. **Davis SS, Gibson DS, Clark R.** 1984. The effect of bovine ephemeral fever on milk production. Aust. Vet. J. **61**:128.
3. **Mellor PS, Boorman J, Baylis M.** 2000. Culicoides biting midges: their role as arbovirus vectors. Annu. Rev. Entomol. **45**:307–340.
4. **Rasmussen LD, Kristensen B, Kirkeby C, Rasmussen TB, Belsham GJ, Bodker R, Botner A.** 2012. Culicoids as vectors of Schmallenberg virus. Emerg. Infect. Dis. **18**:1204–1206.
5. **Brackney DE, Beane JE, Ebel GD.** 2009. RNAi targeting of West Nile virus in mosquito midguts promotes virus diversification. PLoS Pathog. **5**:e1000502. doi:10.1371/journal.ppat.1000502.
6. **Brackney DE, Scott JC, Sagawa F, Woodward JE, Miller NA, Schilkey FD, Mudge J, Wilusz J, Olson KE, Blair CD, Ebel GD.** 2010. C6/36 Aedes albopictus cells have a dysfunctional antiviral RNA interference response. PLoS Negl. Trop. Dis. **4**:e856. doi:10.1371/journal.pntd.0000856.
7. **Campbell CL, Keene KM, Brackney DE, Olson KE, Blair CD, Wilusz J, Foy BD.** 2008. Aedes aegypti uses RNA interference in defense against Sindbis virus infection. BMC Microbiol. **8**:47.
8. **Keene KM, Foy BD, Sanchez-Vargas I, Beaty BJ, Blair CD, Olson KE.** 2004. RNA interference acts as a natural antiviral response to O'nyong-nyong virus (Alphavirus; Togaviridae) infection of Anopheles gambiae. Proc. Natl. Acad. Sci. U. S. A. **101**:17240–17245.
9. **Khoo CC, Piper J, Sanchez-Vargas I, Olson KE, Franz AW.** 2010. The RNA interference pathway affects midgut infection and escape barriers for Sindbis virus in Aedes aegypti. BMC Microbiol. **10**:130.
10. **Sanchez-Vargas I, Scott JC, Poole-Smith BK, Franz AW, Barbosa-Solomieu V, Wilusz J, Olson KE, Blair CD.** 2009. Dengue virus type 2 infections of Aedes aegypti are modulated by the mosquito's RNA interference pathway. PLoS Pathog. **5**:e1000299. doi:10.1371/journal.ppat.1000299.
11. **Scott JC, Brackney DE, Campbell CL, Bondu-Hawkins V, Hjelle B, Ebel GD, Olson KE, Blair CD.** 2010. Comparison of Dengue virus type 2-specific small RNAs from RNA interference-competent and -incompetent mosquito cells. PLoS Negl. Trop. Dis. **4**:e848. doi:10.1371/journal.pntd.0000848.
12. **Siu RW, Fragkoudis R, Simmonds P, Donald CL, Chase-Topping ME, Barry G, Attarzadeh-Yazdi G, Rodriguez-Andres J, Nash AA, Merits A, Fazakerley JK, Kohl A.** 2011. Antiviral RNA interference responses induced by Semliki Forest virus infection of mosquito cells: characterization, origin, and frequency-dependent functions of virus-derived small interfering RNAs. J. Virol. **85**:2907–2917.
13. **Blair CD.** 2011. Mosquito RNAi is the major innate immune pathway controlling arbovirus infection and transmission. Future Microbiol. **6**:265–277.
14. **Donald CL, Kohl A, Schnettler E.** 2012. New insights into control of arbovirus replication and spread by insect RNA interference pathways. Insects **3**:511–531.
15. **Fragkoudis R, Attarzadeh-Yazdi G, Nash AA, Fazakerley JK, Kohl A.** 2009. Advances in dissecting mosquito innate immune responses to arbovirus infection. J. Gen. Virol. **90**:2061–2072.
16. **Bernstein E, Caudy AA, Hammond SM, Hannon GJ.** 2001. Role for a

bidentate ribonuclease in the initiation step of RNA interference. Nature **409**:363–366.
17. **Deddouche S, Matt N, Budd A, Mueller S, Kemp C, Galiana-Arnoux D, Dostert C, Antoniewski C, Hoffmann JA, Imler JL.** 2008. The DExD/H-box helicase Dicer-2 mediates the induction of antiviral activity in drosophila. Nat. Immunol. **9**:1425–1432.
18. **Hess AM, Prasad AN, Ptitsyn A, Ebel GD, Olson KE, Barbacioru C, Monighetti C, Campbell CL.** 2011. Small RNA profiling of Dengue virus-mosquito interactions implicates the PIWI RNA pathway in anti-viral defense. BMC Microbiol. **11**:45.
19. **Morazzani EM, Wiley MR, Murreddu MG, Adelman ZN, Myles KM.** 2012. Production of virus-derived ping-pong-dependent piRNA-like small RNAs in the mosquito soma. PLoS Pathog. **8**:e1002470. doi:10.1371/journal.ppat.1002470.
20. **Myles KM, Wiley MR, Morazzani EM, Adelman ZN.** 2008. Alphavirus-derived small RNAs modulate pathogenesis in disease vector mosquitoes. Proc. Natl. Acad. Sci. U. S. A. **105**:19938–19943.
21. **Ding SW.** 2010. RNA-based antiviral immunity. Nat. Rev. Immunol. **10**:632–644.
22. **Ding SW, Voinnet O.** 2007. Antiviral immunity directed by small RNAs. Cell **130**:413–426.
23. **Kemp C, Imler JL.** 2009. Antiviral immunity in drosophila. Curr. Opin. Immunol. **21**:3–9.
24. **Adelman ZN, Sanchez-Vargas I, Travanty EA, Carlson JO, Beaty BJ, Blair CD, Olson KE.** 2002. RNA silencing of dengue virus type 2 replication in transformed C6/36 mosquito cells transcribing an inverted-repeat RNA derived from the virus genome. J. Virol. **76**:12925–12933.
25. **Attarzadeh-Yazdi G, Fragkoudis R, Chi Y, Siu RW, Ulper L, Barry G, Rodriguez-Andres J, Nash AA, Bouloy M, Merits A, Fazakerley JK, Kohl A.** 2009. Cell-to-cell spread of the RNA interference response suppresses Semliki Forest virus (SFV) infection of mosquito cell cultures and cannot be antagonized by SFV. J. Virol. **83**:5735–5748.
26. **Mellor PSB, Mand Mertens PPC (ed).** 2009. Bluetongue. Biology of animal infections. Academic Press, Amsterdam, Netherlands.
27. **Schwartz-Cornil I, Mertens PP, Contreras V, Hemati B, Pascale F, Breard E, Mellor PS, MacLachlan NJ, Zientara S.** 2008. Bluetongue virus: virology, pathogenesis and immunity. Vet. Res. **39**:46.
28. **Mellor PS.** 1990. The replication of bluetongue virus in Culicoides vectors. Curr. Top. Microbiol. Immunol. **162**:143–161.
29. **Mertens PP, Burroughs JN, Walton A, Wellby MP, Fu H, O'Hara RS, Brookes SM, Mellor PS.** 1996. Enhanced infectivity of modified bluetongue virus particles for two insect cell lines and for two Culicoides vector species. Virology **217**:582–593.
30. **Wechsler SJ, McHolland LE.** 1988. Susceptibilities of 14 cell lines to bluetongue virus infection. J. Clin. Microbiol. **26**:2324–2327.
31. **DuToit RM.** 1944. The transmission of blue-tongue and horse sickness by Culicoides. J. Vet. Sci. An. Industry **19**:7–16.
32. **Mellor PS, Jennings M, Boorman JP.** 1984. Culicoides from Greece in relation to the spread of bluetongue virus. Rev. Elev. Med. Vet. Pays. Trop. **37**:286–289.
33. **Carpenter S, Lunt HL, Arav D, Venter GJ, Mellor PS.** 2006. Oral susceptibility to bluetongue virus of Culicoides (Diptera: Ceratopogonidae) from the United Kingdom. J. Med. Entomol. **43**:73–78.
34. **Savini G, Goffredo M, Monaco F, Di Gennaro A, Cafiero MA, Baldi L, de Santis P, Meiswinkel R, Caporale V.** 2005. Bluetongue virus isolations from midges belonging to the Obsoletus complex (Culicoides, Diptera: Ceratopogonidae) in Italy. Vet. Rec. **157**:133–139.
35. **Tabachnick WJ.** 1996. Culicoides variipennis and bluetongue-virus epidemiology in the United States. Annu. Rev. Entomol. **41**:23–43.
36. **Tabachnick WJ.** 1992. Microgeographic and temporal genetic variation in populations of the bluetongue virus vector Culicoides variipennis (Diptera: Ceratopogonidae). J. Med. Entomol. **29**:384–394.
37. **Gouet P, Diprose JM, Grimes JM, Malby R, Burroughs JN, Zientara S, Stuart DI, Mertens PP.** 1999. The highly ordered double-stranded RNA genome of bluetongue virus revealed by crystallography. Cell **97**:481–490.
38. **Grimes JM, Burroughs JN, Gouet P, Diprose JM, Malby R, Zientara S, Mertens PP, Stuart DI.** 1998. The atomic structure of the bluetongue virus core. Nature **395**:470–478.
39. **Roy P.** 2008. Functional mapping of bluetongue virus proteins and their interactions with host proteins during virus replication. Cell Biochem. Biophys. **50**:143–157.
40. **Mertens PP, Brown F, Sangar DV.** 1984. Assignment of the genome segments of bluetongue virus type 1 to the proteins which they encode. Virology **135**:207–217.
41. **Ratinier M, Caporale M, Golder M, Franzoni G, Allan K, Nunes SF, Armezzani A, Bayoumy A, Rixon F, Shaw A, Palmarini M.** 2011. Identification and characterization of a novel non-structural protein of bluetongue virus. PLoS Pathog. **7**:e1002477. doi:10.1371/journal.ppat.1002477.
42. **Hoffmann B, Scheuch M, Hoper D, Jungblut R, Holsteg M, Schirrmeier H, Eschbaumer M, Goller KV, Wernike K, Fischer M, Breithaupt A, Mettenleiter TC, Beer M.** 2012. Novel orthobunyavirus in cattle, Europe, 2011. Emerg. Infect. Dis. **18**:469–472.
43. **Kupferschmidt K.** 2012. Infectious disease. Scientists rush to find clues on new animal virus. Science **335**:1028–1029.
44. **Vodovar N, Bronkhorst AW, van Cleef KW, Miesen P, Blanc H, van Rij RP, Saleh MC.** 2012. Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. PLoS One **7**:e30861. doi:10.1371/journal.pone.0030861.
45. **Arnaud F, Black SG, Murphy L, Griffiths DJ, Neil SJ, Spencer TE, Palmarini M.** 2010. Interplay between ovine bone marrow stromal cell antigen 2/tetherin and endogenous retroviruses. J. Virol. **84**:4415–4425.
46. **Wechsler SJ, McHolland LE, Tabachnick WJ.** 1989. Cell lines from Culicoides variipennis (Diptera: Ceratopogonidae) support replication of bluetongue virus. J. Invertebr. Pathol. **54**:385–393.
47. **Dulbecco R, Vogt M.** 1953. Some problems of animal virology as studied by the plaque technique. Cold Spring Harb. Symp. Quant. Biol. **18**:273–279.
48. **Schnettler E, Hemmes H, Goldbach R, Prins M.** 2008. The NS3 protein of rice hoja blanca virus suppresses RNA silencing in mammalian cells. J. Gen. Virol. **89**:336–340.
49. **Reed LJ, Muench H.** 1938. A simple method of estimating fifty percent endpoints. Am. J. Hygiene **27**:493–497.
50. **Varela M, Schnettler E, Caporale M, Murgia C, Barry G, McFarlane M, McGregor E, Piras IM, Shaw A, Lamm C, Janowicz A, Beer M, Glass M, Herder V, Hahn K, Baumgärtner W, Kohl A, Palmarini M.** 2013. Schmallenberg virus pathogenesis, tropism and interaction with the innate immune system of the host. PLoS Pathog. **9**:e1003133. doi:10.1371/journal.ppat.1003133.
51. **Martin M.** 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet **17**:3.
52. **Core Team R.** 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
53. **Bembom O.** seqLogo: sequence logos for DNA sequence alignments, R package version 1.22.0.
54. **Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J.** 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5**:R80.
55. **van Cleef KW, van Mierlo JT, van den Beek M, van Rij RP.** 2011. Identification of viral suppressors of RNAi by a reporter assay in Drosophila S2 cell culture. Methods Mol. Biol. **721**:201–213.
56. **Saleh MC, van Rij RP, Hekele A, Gillis A, Foley E, O'Farrell PH, Andino R.** 2006. The endocytic pathway mediates cell entry of dsRNA to induce RNAi silencing. Nat. Cell Biol. **8**:793–802.
57. **Ulvila J, Parikka M, Kleino A, Sormunen R, Ezekowitz RA, Kocks C, Ramet M.** 2006. Double-stranded RNA is internalized by scavenger receptor-mediated endocytosis in Drosophila S2 cells. J. Biol. Chem. **281**:14370–14375.
58. **Boyce M, Celma CC, Roy P.** 2012. Bluetongue virus non-structural protein 1 is a positive regulator of viral protein synthesis. Virol. J. **9**:178.
59. **Owens RJ, Limn C, Roy P.** 2004. Role of an arbovirus nonstructural protein in cellular pathogenesis and virus release. J. Virol. **78**:6649–6656.
60. **Chotkowski HL, Ciota AT, Jia Y, Puig-Basagoiti F, Kramer LD, Shi PY, Glaser RL.** 2008. West Nile virus infection of Drosophila melanogaster induces a protective RNAi response. Virology **377**:197–206.
61. **Saito K, Siomi MC.** 2010. Small RNA-mediated quiescence of transposable elements in animals. Dev. Cell **19**:687–697.
62. **Senti KA, Brennecke J.** 2010. The piRNA pathway: a fly's perspective on the guardian of the genome. Trends Genet. **26**:499–509.
63. **Siomi MC, Miyoshi T, Siomi H.** 2010. piRNA-mediated silencing in Drosophila germlines. Semin. Cell Dev. Biol. **21**:754–759.

64. **Siomi MC, Sato K, Pezic D, Aravin AA.** 2011. PIWI-interacting small RNAs: the vanguard of genome defence. Nat. Rev. Mol. Cell Biol. **12**:246–258.
65. **van Rij RP, Berezikov E.** 2009. Small RNAs and the control of transposons and viruses in Drosophila. Trends Microbiol. **17**:163–171.
66. **Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ.** 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell **128**:1089–1103.
67. **Shih JD, Hunter CP.** 2011. SID-1 is a dsRNA-selective dsRNA-gated channel. RNA **17**:1057–1065.
68. **Shah PS, Schaffer DV.** 2011. Antiviral RNAi: translating science towards therapeutic success. Pharm. Res. **28**:2966–2982.
69. **Mueller S, Gausson V, Vodovar N, Deddouche S, Troxler L, Perot J, Pfeffer S, Hoffmann JA, Saleh MC, Imler JL.** 2010. RNAi-mediated immunity provides strong protection against the negative-strand RNA vesicular stomatitis virus in Drosophila. Proc. Natl. Acad. Sci. U. S. A. **107**:19390–19395.
70. **Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, Dalmay T.** 2012. Reducing ligation bias of small RNAs in libraries for next generation sequencing. Silence **3**:4.
71. **Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB.** 2012. Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. Nucleic Acids Res. **40**:e54.
72. **Eaton BT, Hyatt AD, Brookes SM.** 1990. The replication of bluetongue virus. Curr. Top. Microbiol. Immunol. **162**:89–118.
73. **Kato H, Sato S, Yoneyama M, Yamamoto M, Uematsu S, Matsui K, Tsujimura T, Takeda K, Fujita T, Takeuchi O, Akira S.** 2005. Cell type-specific involvement of RIG-I in antiviral response. Immunity **23**:19–28.
74. **Marques JT, Devosse T, Wang D, Zamanian-Daryoush M, Serbinowski P, Hartmann R, Fujita T, Behlke MA, Williams BR.** 2006. A structural basis for discriminating between self and nonself double-stranded RNAs in mammalian cells. Nat. Biotechnol. **24**:559–565.
75. **Mertens PP, Diprose J.** 2004. The bluetongue virus core: a nano-scale transcription machine. Virus Res. **101**:29–43.
76. **Roy P.** 2008. Bluetongue virus: dissection of the polymerase complex. J. Gen. Virol. **89**:1789–1804.
77. **Chauveau E, Doceul V, Lara E, Adam M, Breard E, Sailleau C, Viarouge C, Desprat A, Meyer G, Schwartz-Cornil I, Ruscanu S, Charley B, Zientara S, Vitour D.** 2012. Sensing and control of bluetongue virus infection in epithelial cells via RIG-I and MDA5 helicases. J. Virol. **86**:11789–11799.
78. **Wu Q, Luo Y, Lu R, Lau N, Lai EC, Li WX, Ding SW.** 2010. Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. Proc. Natl. Acad. Sci. U. S. A. **107**:1606–1611.
79. **Zambon RA, Vakharia VN, Wu LP.** 2006. RNAi is an antiviral immune response against a dsRNA virus in Drosophila melanogaster. Cell Microbiol. **8**:880–889.
80. **Weber F, Wagner V, Rasmussen SB, Hartmann R, Paludan SR.** 2006. Double-stranded RNA is produced by positive-strand RNA viruses and DNA viruses but not in detectable amounts by negative-strand RNA viruses. J. Virol. **80**:5059–5064.
81. **Leger P, Lara E, Jagla B, Sismeiro O, Mansuroglu Z, Coppee JY, Bonnefoy E, Bouloy M.** 21 November 2012. Dicer-2 and Piwi mediated RNA interference in Rift Valley fever virus infected mosquito cells. J. Virol. [Epub ahead of print.] doi:10.1128/JVI.02795-12.
82. **Li F, Ding SW.** 2006. Virus counterdefense: diverse strategies for evading the RNA-silencing immunity. Annu. Rev. Microbiol. **60**:503–531.
83. **Schnettler E, Sterken MG, Leung JY, Metz SW, Geertsema C, Goldbach RW, Vlak JM, Kohl A, Khromykh AA, Pijlman GP.** 2012. Noncoding flavivirus RNA displays RNA interference suppressor activity in insect and mammalian cells. J. Virol. **86**:13486–13500.

9436–9446 *Nucleic Acids Research, 2014, Vol. 42, No. 14* *Published online 22 July 2014*
doi: 10.1093/nar/gku657

Induction and suppression of tick cell antiviral RNAi responses by tick-borne flaviviruses

Esther Schnettler[1,2,*], Hana Tykalová[3], Mick Watson[2], Mayuri Sharma[4], Mark G. Sterken[5], Darren J. Obbard[6], Samuel H. Lewis[6], Melanie McFarlane[1], Lesley Bell-Sakyi[2], Gerald Barry[2], Sabine Weisheit[2], Sonja M. Best[7], Richard J. Kuhn[4], Gorben P. Pijlman[5], Margo E. Chase-Topping[8], Ernest A. Gould[9,10], Libor Grubhoffer[3], John K. Fazakerley[2] and Alain Kohl[1,2,*]

[1]MRC - University of Glasgow Centre for Virus Research, Glasgow G11 5JR, UK, [2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK, [3]Faculty of Science, University of South Bohemia and Biology Centre, Institute of Parasitology, Czech Academy of Sciences, 37005 České Budějovice (Budweis), Czech Republic, [4]Markey Centre for Structural Biology, Department of Biological Sciences, Purdue University, West Lafayette IN 47907, USA, [5]Laboratory of Virology, Wageningen University, 6708 PB Wageningen, The Netherlands, [6]Institute of Evolutionary Biology and Centre for Infection Immunity and Evolution, University of Edinburgh, EH9 3JT, UK, [7]Innate Immunity and Pathogenesis Unit, Laboratory of Virology, Rocky Mountain Laboratories, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT 59840, USA, [8]Centre for Immunity, Infection and Evolution, University of Edinburgh, EH9 3JT, UK, [9]Unité des Virus Emergents, Faculté de Médicine Timone, 13385 Marseille Cedex 05, France and [10]Centre for Hydrology and Ecology, Maclean Building, Oxon OX10 8BB, UK

Received December 20, 2013; Revised July 4, 2014; Accepted July 8, 2014

ABSTRACT

Arboviruses are transmitted by distantly related arthropod vectors such as mosquitoes (class *Insecta*) and ticks (class *Arachnida*). RNA interference (RNAi) is the major antiviral mechanism in arthropods against arboviruses. Unlike in mosquitoes, tick antiviral RNAi is not understood, although this information is important to compare arbovirus/host interactions in different classes of arbovirus vectos. Using an *Ixodes scapularis*-derived cell line, key Argonaute proteins involved in RNAi and the response against tick-borne Langat virus (*Flaviviridae*) replication were identified and phylogenetic relationships characterized. Analysis of small RNAs in infected cells showed the production of virus-derived small interfering RNAs (viRNAs), which are key molecules of the antiviral RNAi response. Importantly, viRNAs were longer (22 nucleotides) than those from other arbovirus vectors and mapped at highest frequency to the termini of the viral genome, as opposed to mosquito-borne flaviviruses. Moreover, tick-borne flaviviruses expressed subgenomic flavivirus RNAs that interfere with tick RNAi. Our results characterize the antiviral RNAi response in tick cells including phylogenetic analysis of genes encoding antiviral proteins, and viral interference with this pathway. This shows important differences in antiviral RNAi between the two major classes of arbovirus vectors, and our data broadens our understanding of arthropod antiviral RNAi.

INTRODUCTION

Tick-borne arboviruses of the *Flaviviridae* family are highly relevant to public health (1). Much work on tick-borne arboviruses has been carried out with Langat virus (LGTV), isolated from *Ixodes granulatus* and *Haemaphysalis* ssp.

*To whom correspondence should be addressed. Tel: +44(0)141 330 3921; Email: alain.kohl@glasgow.ac.uk
Correspondence may also be addressed to Esther Schnettler. Tel: +44(0)141 330 0233; Fax: +44(0)141 330 3520; Esther.Schnettler@glasgow.ac.uk
Present addresses:
Mayuri Sharma, Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA.
John K. Fazakerley, Sabine Weisheit & Lesley Bell-Sakyi, The Pirbright Institute, Ash Road, Pirbright, Surrey GU24 0NF, UK.
Gerald Barry, MRC—University of Glasgow Centre for Virus Research, Glasgow G11 5JR, UK.

ticks in Malaysia and Thailand and related to tick-borne encephalitis virus (TBEV) (1–4). Flaviviruses are positive-stranded RNA viruses. Viral proteins are encoded in a single open reading frame. The untranslated RNA regions (UTRs) at the genome termini regulate replication and translation (5–8).

Arbovirus infection of arthropod cells is characterized by little or no cytopathic effects (9). Studies of vector/arbovirus interactions suggests that this may be at least partly due to regulation of arbovirus replication by innate immune responses (10). Research on vector immune responses to arboviruses has focused on mosquitoes (11,12) despite the fact that many European/Asian arboviruses are tick-borne (13). Antiviral responses in mosquitoes rely on a small RNA-based mechanism called RNA interference (RNAi) (10,11). The exogenous small interfering (si)RNA pathway is especially important and can be induced by virus-derived long double-stranded (ds)RNA molecules generated during infection (either replication intermediates or secondary RNA structures) or dsRNA viral genome (10). In insects, dsRNA is targeted by the Dicer enzyme (Dcr-2) and cleaved into 21 nucleotide (nt) siRNAs, also known as viRNAs (10,11). In *Drosophila*, viRNAs are integrated into the Argonaute-2 protein (Ago-2) containing RNA-induced silencing complex, unwound and one strand of the viRNA is retained by Ago-2 to guide degradation of complementary (viral) RNA (14). Other Ago and Dcr proteins, i.e. Dcr-1 and Ago-1, are involved in the microRNA (miRNA) pathway (10–11,14).

Following treatment with gene-specific dsRNA or siRNAs, ticks and tick cell cultures can induce sequence-specific RNAi of endogenous genes (15) and restrict viral infections (16–18). Sequence analysis has also identified putative Ago and Dcr genes in the *I. scapularis* genome (19). However, it is not known if these are transcribed and involved in tick antiviral RNAi responses. All studied insect specific viruses and plant-infecting viruses have been shown to express RNA silencing suppressor (RSS) proteins which interfere with the RNAi response (20). No RSS proteins have been identified for arboviruses although evasion strategies have been suggested for the alphavirus Semliki Forest virus (SFV) (21), and the production of a subgenomic flavivirus RNA (sfRNA) interfering with the RNAi response was reported for mosquito-borne flaviviruses (22).

In this study, we identify and characterize key RNAi players of the Ago family that interfere with LGTV replication and describe characteristics of viRNAs in tick vector cells, which are different to viRNAs in mosquitoes. We also demonstrate that the recently described RSS activity of mosquito-borne flavivirus sfRNA can be broadened to tick-borne LGTV and TBEV sfRNA. The results imply that the antiviral RNAi system in ticks is more complex and has important differences to that of mosquitoes.

MATERIALS AND METHODS

Viruses and plasmids

The LGTV replicon (E5repRluc2B/3) was derived from the infectious cDNA of LGTV E5 (4). Modifications in the LGTV replicon were based on the previously described replicon construct for TBEV Neudoerfl strain (23). This construct encodes the first 17 residues of capsid, followed by the Rluc gene, the last 27 residues of the envelope and all non-structural proteins, as described in Supplementary Data. For infections of tick cells, LGTV strain TP21 was used.

Invertebrate expression vectors, pIZ-Fluc, pAcIE1-Rluc and pIB-MBP-HDVr have been described previously (22). The 3′UTRs of LGTV and TBEV were amplified by polymerase chain reaction (PCR) using, respectively, E5repRluc2B/3 or pTNd/ΔME (24) as templates. Invertebrate expression plasmids were obtained by fusing the 3′ terminus to the HDVr sequence from a WNV 3′UTR expression construct (22) using PCR. The resulting products were cloned into pDonor207 and pIB-GW plasmids (Invitrogen) using Gateway technology.

Luciferase assays

Luciferase activities were determined using a Dual Luciferase assay kit (Promega) in a GloMax multi-luminometer following cell lysis in Passive Lysis Buffer.

Cell culture, transfection and infection

BHK-21 cells were grown in GMEM at 37°C as previously described (25). Cells (3×10^5/well) were seeded in a 6-well plate prior to transfection with Lipofectamine2000 (Invitrogen) according to the manufacture's protocol. The *I. scapularis*-derived IDE8 cells were grown in L-15B medium (26) at 32°C in ambient air as previously described (27). Cells (6.5×10^5/well) were seeded in 24-well plates prior to transfection. Transient RNAi suppression assays were performed by transfecting 200 ng pIZ-Fluc, 300 ng pAcIE1-Rluc and 500 ng pIB-MBP-HDVr, TBEV 3′UTR or LGTV 3′UTR into IDE8 cells using Genejammer (Agilent) following the manufacturer's instructions. Silencing of reporter genes was induced at 24 h post-transfection (hpt) through addition of 280 ng dsRNA to the cell culture medium; luciferase was measured 48 hpt.

In case of studies involving replicon, putative RNAi genes were silenced by the addition of 300 ng dsRNA to cell culture medium at 6 and 30 h post-seeding (hps). Then, capped *in vitro*-transcribed E5repRluc2B/3 was transfected 48 hps using Lipofectamine2000 according to the manufacturer's instructions. Luciferase expression was measured 24 hpt.

For infection assays, target genes were first silenced by transfection of 100 ng dsRNA using Lipofectamine2000, followed by LGTV TP21 infection at 24 hpt at a multiplicity of infection (MOI) of 0.1. RNA was isolated at 48 h post infection (hpi) by Trizol.

Statistical analysis

The relative luciferase expression (RL) was calculated as:

$$RL_i = I_{Fluc,i}/I_{Rluc,i}$$

Where I is the measured intensity and i is the sample. To cancel out construct specific effects, values under treatment (for example co-transfected with dsFFluc) were normalized

against the same construct that was treated with a negative control (in this example dseGFP). Thus:

$$\mathrm{NRL_x} = \mathrm{RL}_{i,\mathrm{treated}}/\mathrm{RL}_{i,\mathrm{neg.control}}$$

Experiments were performed in duplicate or in triplicate and repeated independently at least three times. The independent experiments were averaged:

$$\overline{\mathrm{NRL}} = \sum_{x}^{n} \frac{\mathrm{NRL}_x}{n}$$

Where x is the xth experiment and n is the total number of experiments.

The significances were calculated using custom-written scripts in R (www.r-project.org). In case of pairwise testing a two-sample independent t-test was performed, as provided by R.

Multiple testing was done by applying Tukey's HSD (also known as Tukey's range test), the q-value was calculated and compared to the indexed q in the studentized range distribution available in R. Significant differences ($P \leq 0.05$) are indicated in the graphs with an *.

Small RNA isolation and deep sequencing analysis

1.5×10^6 cells per tube were either transfected with 1 μg of eGFP-derived dsRNA, capped *in vitro*-transcribed E5repRluc2B/3 RNA, infected with LGTV TP21 (MOI 10) or untreated. At 48 hpt or 72 hpi, RNA was isolated using 1 ml Trizol (Invitrogen) per tube, small RNAs of 18–30 nt were sequenced and analyzed using viRome as previously described (28,29). Small RNA data was submitted to the European Nucleotide Archive (accession number ERP006219).

Reverse transcription and PCR

RNA was isolated by Trizol, following the manufacturer's protocol. Total RNA (500 ng for untreated/dsRNA-treated cells as well as knockdowns followed by LGTV infection or 5 μg LGTV antigenome detection) was reverse transcribed with Superscript III (Invitrogen) and using either oligo dT primers (knockdowns), an antigenome specific primer (LGTV antigenome detection) or random hexamers (LGTV infection) following the manufacturer's instructions. For the detection and amplification of Ago and Dcr transcripts, PCR was carried out using 2 μl of the cDNA reaction with corresponding primers (Table 1). The eGFP-derived PCR product was produced using eGFP-C1 (Clontech) as template. In case of LGTV antigenome detection, two rounds of PCR were performed using LGTV specific primers. PCR products were gel-purified, cloned into the pJet blunt1.2 vector (Fermentas) and sequenced.

LGTV RNA was determined by QRT-PCR with NS5 specific primers using the Fast SYBR Green PCR Master Mix (Life Technologies) according to manufacturer's instructions. Previously described actin primers were used as housekeeping genes (16).

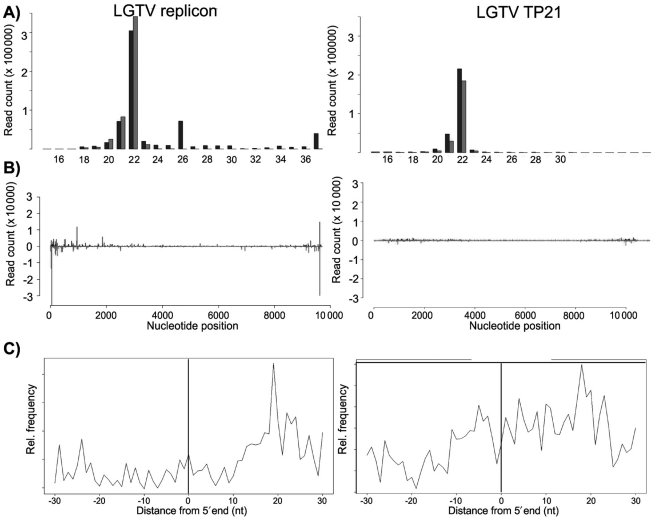


Figure 1. Characterization of exogenous-derived small RNAs in IDE8 cells. (**A**) Size distribution of small RNA molecules mapping either to LGTV E5repRluc2B/3 replicon (left panel) at 48 hpt or LGTV TP21 (right panel) at 72 hpi in IDE8 cells. (**B**) Frequency distribution of 22 nt small RNA molecules mapped to the E5repRluc 2B/3 replicon (5′UTR to 3′UTR) (left panel) or LGTV TP21 (right panel). The y-axis shows the frequency of the 22 nt siRNAs mapping to the corresponding nucleotide position in the x-axis. Positive numbers and dark gray peaks represent the frequency of siRNAs mapping to the genome (in 5′-3′ orientation) and light gray peaks/negative numbers to the antigenome (in 3′-5′ orientation). See also Supplementary Figure S2. (**C**) Frequency map of 22 nt small RNAs mapping to the opposite strand of the LGTV replicon (left panel) or LGTV TP21 (right panel).

In vitro transcription and dsRNA production

E5repRluc2B/3 was linearized by EcoRV and *in vitro*-transcribed using a SP6 Megascript kit (Ambion) in the presence of cap analogue according to the manufacturer's protocol. dsRNA was produced with the RNAi Megascript kit (Ambion) from PCR products flanked by T7 promoter sequences.

Phylogenetic analysis

To place the *Ixodes* sequences within gene trees, representative sequences were downloaded from Genbank (see Supplementary Figure S5 for sequence identifiers) for selected arthropods (waterfleas, copepods, lice, ticks, centipedes, flies, butterflies, beetles and wasps) and deuterostomes (sea squirt, human, chicken and zebrafish) that have sufficient complete genomes and/or transcriptomes. Ago and Piwi were aligned with translational MAFFT (30) and poorly-aligned regions were removed manually, resulting in an aligned matrix of 2349 positions for Ago and 2241 positions for Piwi. Due to a higher level of sequence divergence and higher proporation of incomplete orthologus sequences, Dicer was aligned under a codon model using PRANK (31), and then GBLOCKs (32) was used to exclude regions of poor alignment, resulting in an aligned matrix of 810 positions. Gene trees were inferred with MrBayes (33) using unlinked General Time Reversible models with Gamma-distributed rate variation for each of the three codon positions. Two parallel MCMC chains of >25 mil-

Table 1. List of primer sequences used

Gene	Upstream/downstream primer sequences (5′-3′)
Ago-68	*gtaatacgactcactatagggCGAGACTTTCAGAGCGTG/* *gtaatacgactcactatagggGTTGGTGTACTTCGCCAT*
Ago-30	*gtaatacgactcactatagggACATACGAGCACTGACGG/* *gtaatacgactcactatagggTGGTGCAACATTTTATCGA*
Ago-30–2	*gtaatacgactcactatagggGAACGCCAAAAAGATCCCA/* *gtaatacgactcactatagggCCGGTACCATCCTCATTTCT*
Ago-16	*gtaatacgactcactatagggAAGATCACGAGGGTATCGGTAGT/* *gtaatacgactcactatagggACTTTTCTGCACCACGTCTTG*
Ago-16-2 (RT-PCR detection)	*gtaatacgactcactatagggCGTTATGAAGGGTGATCAGAAG/* *gtaatacgactcactatagggGACTGGTACTGATTCTCCCA*
Ago-96	*gtaatacgactcactatagggATGCCTGCTCGGACATCTAC/ gtaata* *cgactcactatagggTCGAGTGAACGTCCAAATTCT*
Ago-78	*gtaatacgactcactatagggGAGGTGAAGCGTGTGGGG/ gtaatacga* *ctcactatagggGATGGAAGGCTTCTTGTTGTC*
Dcr-90	*gtaatacgactcactatagggATCCTCAAGGAGTACAAGCC/ gtaata* *cgactcactatagggACAGAGCATTAGGGTCGTC*
Dcr-98	*gtaatacgactcactatagggATCCCGTCTTTCCCGATCTT/* *gtaatacgactcactatagggTGCATCACAGGTGCCAGG*
eGFP	*gtaatacgactcactatagggGGCGTGCAGTGCTTCAGCCGC/* *gtaatacgactcactatagggGTGGTTGTCGGGCAGCAGCAC*
Firefly luciferase	*gtaatacgactcactatagggATGGAAGCAGCCAAAAAC/* *gtaatacgactcactatagggTTACACGGCGATCTTTCC*
LGTV antigenome RT-PCR	aattccacccatgaaatgtac
LGTV NS5 (QT-PCR)	acccaagactgctacgtgtggaaa/tgaggaagtaaagggccttgctga

T7 promoter region is indicated in italics.

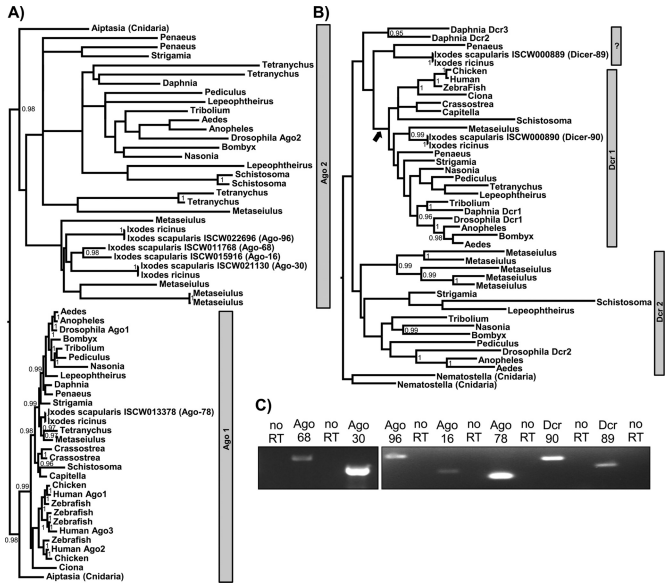


Figure 2. Analysis of Ago and Dcr protein-encoding genes in the *Ixodes scapularis* genome. (**A**) and (**B**) are gene trees for metazoan Ago-subfamily genes **(A)**and Dcr genes **(B)** respectively, constructed using a Bayesian approach under a GTR model (nodes are labeled if they receive >90% support; see Materials and Methods). Trees are unrooted, but presented as if the root fell between the two Cnidarian homologs. (**C**) Detection of transcripts encoding Ago and Dcr proteins in IDE8 cells by RT-PCR. RNA not treated with reverse transcriptase was used as control during the PCR reaction (no RT). See also Supplementary Figure S5.

lion steps were run for each tree (sampling every 1000 steps) and the first 25% of steps were discarded as burn-in. Stationarity was inferred by comparing parallel runs and inspection of the chains for each parameter: Potential Scale Reduction Factors approached 1.000, the variation in split frequencies was <0.01, and effective sample sizes were >1000 for all parameters. The trees presented in Figure 2 and Supplementary Figure S5 are maximum clade credibility trees, with branch lengths proportional to the number of substitutions. Only partitions with >90% Bayesian Posterior support are labeled.

RNA structure predictions

Consensus RNA structures were predicted using the LocARNA web server (Vienna RNA web server 1.8.2) (34) with standard settings. Pseudoknots were identified manually. Thermodynamic stability was calculated by folding an individual sequence with RNAfold (Vienna RNA web server 1.8.2), using a secondary structure constraint and standard settings.

Northern blot analysis

Northern blot analysis was performed by loading 4.5 μg or 3 μg of total RNA of BHK-21 or tick cells, respectively, on a 1.5% agarose-2% formaldehyde MOPS gel and transferred to a nitrocellulose (Hybond-N+, GE Healthcare) membrane using 'top down' blotting with 20xSSC as transfer buffer. Transferred RNA was UV-crosslinked for 2 min. Hybridization was performed for 2 h in HybPerfect buffer (Sigma) at 63°C using DIG-labeled PCR product as probe (TBEV 3′UTR or LGTV 3′ UTR). Membranes were washed twice with 2xSSC + 0.1% SDS for 5 min, twice with 0.2xSSC + 0.1% SDS for 20 min at 63°C and DIG was detected using an anti-DIG antibody as described previously (35).

RESULTS

scapularis-derived-IDE8 cells mount RNAi responses against LGTV and TBEV

An uncharacterized RNAi response was shown to restrict mosquito-borne arbovirus infections in *I. scapularis*-derived ISE6 and IDE8 cells (16–18). It is not known if RNAi is induced in tick cells following infection with tick-borne arboviruses. Production of viRNAs is an indicator of an antiviral RNAi response. An LGTV E5 strain replicon encoding the *Renilla* luciferase (Rluc) gene as a reporter (E5repRluc2B/3) was constructed to investigate antiviral RNAi in IDE8 cells (Supplementary Figure S1A). The ability to successfully transfect E5repRluc2B/3 RNA into IDE8 cells (77%) was determined, using either fluorescently labeled replicon RNA or immune-fluorescence detection of NS3, respectively (Supplementary Figure S6A). Following transfection with E5repRluc2B/3 RNA, IDE8 cells were lysed and Rluc expression determined at 24, 48, 72, 96 and 120 hpt. Expression was observed 24 hpt then decreased (Supplementary Figure S1C). Replication was verified by detection of LGTV antigenome (Supplementary Figure S1B). These results suggest that the LGTV replicon is inhibited by an induced antiviral response in IDE8 cells.

Previous work has documented the production of viRNAs in ISE6 cells; however, the sequences and their distribution on the virus genome are not known (17,18). The production of LGTV-specific viRNAs in IDE8 cells was therefore analyzed. At 48 hpt, total RNA was isolated and small RNAs sequenced; frequencies and LGTV genome location of small RNAs were determined (Table 2). 7.1% of the small RNA sequences mapped to the LGTV replicon sequence. viRNAs were predominantly 22 nt in length (59.6%) and mapped with similar frequency to the genome and antigenome (Figure 1A, left panel). viRNAs were scattered along the LGTV replicon genome/antigenome with variable frequency into hot spots/cold spots (21) (Figure 1B, left panel). The 5′ and 3′UTRs generated the highest viRNA frequencies (Figure 1B, left panel). Comparing the base composition of 22 nt viRNAs of hot spots versus cold spots showed a substantial bias away from G toward A at the 5′ end ($P < 0.0001$, Fishers exact test [FET]) and a bias away from A at the 3′ end ($P < 0.0001$, FET). Bias at other positions was found but none was particularly striking (Supplementary Figure S2A). The 5′ ends of the complementary LGTV specific 22 nt RNAs were most frequently separated by 20 nt (Figure 1C, left panel) suggesting generation from dsRNA of 20 nt with 2 nt overhangs. Experiments performed with a previously described TBEV replicon (23) (Supplementary Figure S2C) showed similar results regarding the predominance of 22 nt viRNAs and 5.9% of total small RNAs mapping to the TBEV replicon (Table 2), with similar frequency to the genome and antigenome. The 22 nt small RNAs mapping to TBEV, are scattered along the genome/-antigenome. Again the highest frequency of viRNAs was generated from the 5′ and 3′ UTRs (Supplementary Figure S2D). Experiments with IDE8 cells infected with LGTV TP21 showed the production of virus specific small RNAs sharing several of the characteristics of LGTV replicon-derived viRNAs, although at a lower overall frequency (0.27% for virus and 7.12% for replicon (Table 2)). The majority of viRNAs were 22 nts in length, most frequently separated by 20 nts and the highest viRNA frequencies were generated from and around the 5′ and 3′ ends of the viral genome/antigenome (Figure 1, right panels and Table 2).

The length of small RNAs in IDE8 cells is a host property

Recent studies have shown that insect viRNAs are generally 21 nt in length, in contrast to nematode *Caenorhabditis elegans* viRNAs of predominantly 22 or 23 nt depending on the virus (21,28,36–47). To determine whether generation of 22 nt as the dominant viRNA length was a property of the cells or the virus, an eGFP-derived dsRNA was transfected into IDE8 cells and small RNAs analyzed. Again, 22 nt was the dominant length (Supplementary Figure S2E) and small RNAs mapped in hot/cold spots along the whole eGFP sequence and its complement (Supplementary Figure S2E).

We also analyzed viRNAs targeting the dsRNA orbivirus St. Croix River virus (SCRV) (48,49), which persistently infects IDE8 cells (Table 2). Again, the majority of SCRV viRNAs were 22 nt, with similar frequencies being detected on the (+) and the (–) strand (Supplementary Figure S3).

To determine the properties of endogenous small RNA molecules such as miRNAs, endogenous siRNAs and PIWI-interacting (pi)RNAs (10) in IDE8 cells, the small RNA profiles from uninfected and treated (eGFP dsRNA and LGTV replicon) IDE8 cells were analyzed. Small RNAs mapping to the *I. scapularis* genome (https://www.vectorbase.org) had a predominant length of 22 nt (44.4%) in all samples, with slightly higher frequencies for the sense orientation. Moreover, a class of small RNA molecules of 27 to 29 nt was identified with a peak at 28 nt (27 nt: 6.7%, 28 nt: 10.1% and 29 nt: 5%) as strongly represented as 21 nt small RNAs (12.5%) (Supplementary Figure S4). This indicates that 22 nt is the dominant length of small RNAs (endogenous or viral) in IDE8 cells.

Identification of Dcr and Ago proteins involved in antiviral RNAi in tick cells

Ago-2 and Dcr-2 proteins are key effectors in the insect antiviral RNAi pathway (10,50). Dcr-1 and Ago-1 are known to be important for the insect miRNA pathway (10,14). Previous sequence analysis has shown that the *I. scapularis* genome contains at least one putative Dcr gene, Dcr-89 (ISCW000889) and two putative Ago subfamily genes; Ago-68 (ISCW011768), Ago-30 (ISCW0021130) (19). In the present study, Basic Local Alignment Search Tool (BLAST) similarity searches with Dcr (Dcr-1 and Dcr-2) and Ago subfamily genes (Ago-1 and Ago-2) of *Drosophila melanogaster* and *Aedes aegypti* were performed to identify further putative homologs in the *I. scapularis* genome. Three additional putative Ago subfamily genes; Ago-96 (ISCW022696), Ago-16 (ISCW015916), Ago-78 (ISCW013378) and another putative Dcr gene, Dcr-90 (ISCW0008890) were identified.

To understand the function of *Ixodes* Ago and Dcr proteins within the wider context of their evolution, gene trees were constructed using a Bayesian approach (Figure 2A and

Table 2. Number of small RNA reads

	Langat virus replicon	St. Croix River virus	Langat virus TP21	TBEV replicon	TBEV NS5 GAA replicon
Genome/coding strand reads	719782	1462846	294390	3906753	3153338
Anti-genome/coding strand reads	553286	1242195	227753	3606509	2921677
Total viral reads	1273068	2705041	522143	7513262	6075015
Reads in total	17875799	18806256	190908946	127799016	127066875

Indicated are read numbers of small RNAs mapping to the genomes and antigenomes of St. Croix River virus, Langat virus replicon, Langat virus TP21 and TBEV replicons.

B; Supplementary Figure S5). The last common ancestor of each of these gene families probably pre-dates the origin of the animals (51), so that saturation and long-branch artifacts make reliable tree inference extremely challenging. Rooting the Ago tree between the two cnidarian paralogs identified two well-supported clades: the slowly evolving clade homologous to drosophila Ago-1 (miRNA pathway) and the rapidly evolving clade homologous to drosophila Ago-2 (siRNA pathway). The Ago-1 clade exactly mirrors the known phylogeny of the species, and clearly identifies Ixodes Ago-78 as an ortholog of drosophila Ago-1. The other four *Ixodes* Ago (-96, -68, -16, -30) then appear as more recent Ago-2 paralogs that have evolved since the last common ancestor of *Arachnida* and Pancrustacea, although a lack of support within this clade makes it hard to draw conclusions beyond this. The Dcr gene-tree also lacks support, and when similarly rooted using the cnidarian paralogs results in a pattern that is hard to interpret. With this rooting, *Ixodes* Dcr-90 clusters with other arachnid Dicers basal to an arthropod clade that includes drosophila Dcr-1, suggesting that Dcr-90 is a Dcr-1 homologue. However, the basal position of *Ixodes* Dcr-89 and the remaining crustacean Dicers is difficult to reconcile with the known organismal phylogeny. If the divergent Cnidarian outgroup is excluded, then an alternative rooting immediately basal to the deuterostome/arthropod Dcr-1 clade (marked by a black arrow in Figure 2B) would place Dcr-89 and the remaining Crustacean Dcrs as the most basally-branching arthropod Dcr-2, consistent with the species phylogeny and suggesting it is homologous to drosophila Dcr-2. Transcription of putative Ago and Dcr genes was verified in IDE8 cells (Figure 2C).

In order to investigate mediators of antiviral activity in IDE8 cells, transcripts of individual Dcr or Ago genes were knocked down by RNAi as previously described (16) and the effect on the LGTV replicon determined. Efficiency of knockdown/silencing of cells treated with dsRNA specific for Ago (Ago-68, Ago-30, Ago-16, Ago-96 and Ago-78) and Dcr (Dcr-90 and Dcr-89) genes was determined by semi-quantitative RT-PCR and quantified in relation to control dsRNA using 16S as loading control (Figure 3). Cells treated with dsRNA against Ago-68, Ago-30, Ago-96, Ago-16, Ago-78 or Dcr-90 showed reduction in target transcript levels (9–40%). No significant reduction of Dcr-89 transcript was observed, due to a high variability between samples (Figure 3).

Following successful individual knockdowns of most putative RNAi genes, the experiment was repeated, LGTV

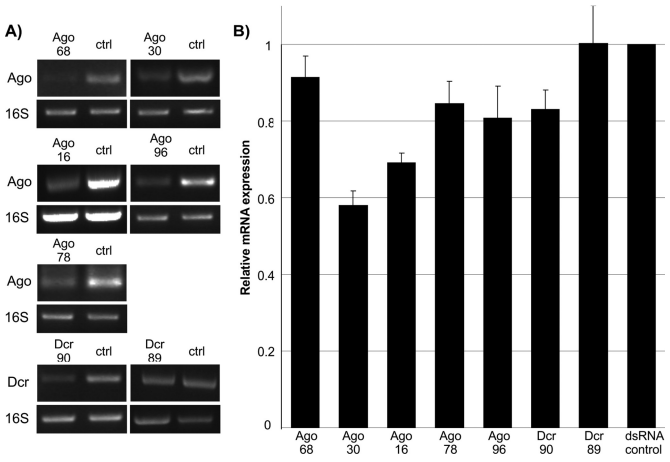


Figure 3. Knockdown of transcripts encoding Ago or Dcr proteins. (**A**) dsRNA-based silencing of Ago and Dcr encoding transcripts in IDE8 cells. Transcripts were detected by RT-PCR using gene-specific primers. A PCR product of 16S ribosomal RNA was used as housekeeping gene and eGFP specific dsRNA treated cells as control (crtl). (**B**) mRNA knockdowns quantification by Image J software, using 16S as control. The relative mean (normalized to eGFP-dsRNA controls) with standard error is shown for at least 10 repeats.

replicon RNA was transfected into silenced IDE8 cells and replicon-mediated Rluc activities determined. Significant increases in replicon Rluc activity were observed in IDE8 cells treated with dsRNA specific for Ago-30, Ago-16 and Dcr-90, compared to control dsRNA (Figure 4A). No significant increase of Rluc was observed following Ago-68, -78, -96 and Dcr-89 knockdowns. To ensure that the observed effect was not due to off target effects, the Ago-30 knockdown was repeated with an additional Ago-30 specific dsRNA molecule (Ago30–2); this resulted in a similar increase of luciferase activity thus confirming previous results (Supplementary Figure S6B). Similar experiments were also performed with silenced IDE8 cells and the effect on LGTV infection (MOI 0.1) at 48 hpi was determined by QRT-PCR. Significant increases in LGTV RNA levels were observed in cells treated with dsRNA specific for Ago-68, -30, -16 and Dcr-89, although Dcr-89 resulted only in a small increase (Figure 4B).

Targeting of the same cells by dsRNA and LGTV replicon or LGTV infection was established, using fluorescently labeled dsRNA and immunostaining of LGTV NS3 or E protein (Supplementary Figure S6A). In summary, tick

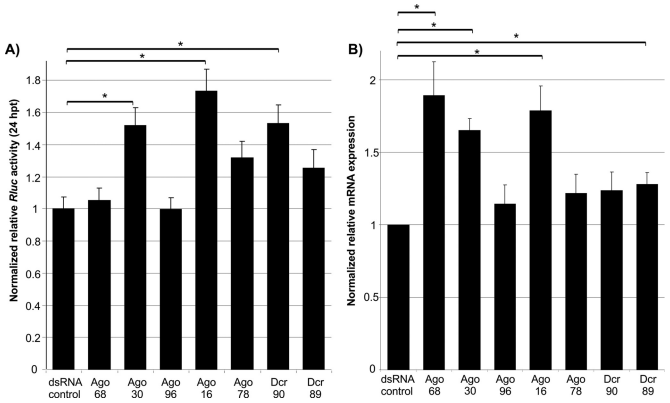


Figure 4. Effects of Ago and Dcr knockdowns on LGTV replication. (**A**) Ago or Dcr silenced cells were transfected with capped *in vitro*- transcribed LGTV E5repRluc2B/3 replicon RNA, and Rluc activity was determined at 24 hpt. The mean with standard error is shown for three independent experiments performed in duplicate (two experiments)/triplicate (one experiment). The data were normalized to cells treated with eGFP-specific control dsRNA. * indicate significance by Tukey's HSD ($P \leq 0.05$). (**B**) Silenced cells were infected with Langat virus (MOI 0.1) and viral RNA was determined by QRT-PCR at 48 hpi, using actin as housekeeping gene internal standard. The mean with standard error is shown for three independent experiments performed in triplicate. The data were normalized to cells treated with eGFP-specific control dsRNA. * indicate significance by Student *t*-test ($P \leq 0.05$).

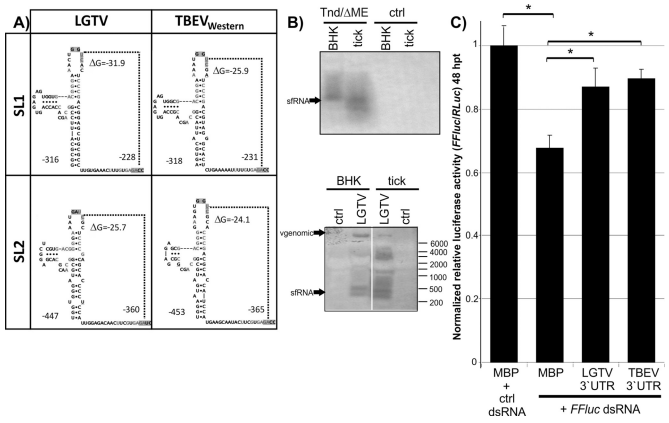


Figure 5. Analysis of subgenomic flavivirus (sf)RNA in the 3′UTR of tick-borne flaviviruses. (**A**) Structure model of SL 2 and SL 1 RNA stem loop structures of TBEV and LGTV. (**B**) Expression of TBEV (TND/ΔME) (top) and LGTV (bottom) sfRNA in replicon (top), non-transfected (control CTRL) or infected cells (bottom) was detected by northern blot analysis with 3′UTR specific DIG-PCR probes. (**C**) The effect of sfRNA on RNAi in IDE8 cells was determined by co-transfection of *FFluc*, Rluc and expression constructs for MBP-HdVr (MBP), LGTV 3′UTR or TBEV 3′UTR. Silencing was induced 24 hpt following addition of dsRNA to the culture medium. At 48 hpt, relative luciferase activity (*FFluc*/Rluc) was determined and normalized to cells treated with eGFP specific (ctrl) dsRNA. The luciferase expression level measured with MBP-HdVr was set at 1.0. The mean with standard error is shown for three independent experiments performed in duplicate (one experiment)/triplicate (two experiments). ∗ indicate significance by Tukey's HSD ($P \leq 0.05$). See also Supplementary Figure S7.

Ago-30 and Ago-16 mediate antiviral activity against both LGTV and its replicon.

Tick-borne subgenomic flavivirus (sf)RNA interferes with antiviral RNAi

sfRNA is derived from the flavivirus 3′UTR, produced in vertebrate and invertebrate cells by mosquito and tick borne-flaviviruses and contains a complex RNA structure (52–54). West Nile virus (WNV) and dengue virus (DENV) sfRNAs both interfere with RNAi (22).

Production of sfRNA and suppression of RNAi by both LGTV and TBEV was investigated. The 3′UTRs of flaviviruses share common characteristics in their RNA architecture (55). It has been demonstrated that mosquito-borne flaviviruses share an RNA stem loop structure (called SL II) toward the 5′ end of the 3′UTR which has similarities to SL IV of the 3′UTR and is important for sfRNA production (52–54). RNA folding predictions of the 3′UTR of tick-borne flaviviruses showed RNA structures with folds highly similar to SL II and SL IV (respectively named SL 2 and SL 1 in the tick-borne viruses) for most tick-borne flaviviruses, despite sequence differences to mosquito-borne flavivirus 3′UTRs (Supplementary Figure S7).

To determine if the predicted LGTV and TBEV RNA stem loop structures (Figure 5A and Supplementary Figure S7) give rise to sfRNAs, vertebrate and tick cells were infected with LGTV (56) or transfected with TBEV replicon (24). Northern blot analysis detected TBEV and LGTV RNA at the expected size of ∼0.4 kb (predicted SL 2, LGTV: −447 nt; TBEV: −453 nt) (Figure 5A and B). In addition, similar to WNV a lower band was observed. This may be due to the presence and characteristics of two SL structures [SL 1 and 2]. Moreover, there are differences between arthropod and vertebrate cells (Figure 5B) (52,53).

The RSS activity of these sfRNAs was investigated next, after establishing successful plasmid transfections in IDE8 cells (Supplementary Figure S6A). IDE8 cells were co-transfected with plasmids encoding Firefly luciferase (*FFluc*; reporter gene), Rluc (internal control), and plasmids expressing LGTV or TBEV 3′UTRs. Maltose binding protein (MBP) sequence fused to the hepatitis delta virus ribozyme (HDVr) was used as negative control RNA as the 3′UTRs plasmids also contain an HDVr. Subsequently, silencing was induced by either *FFluc*-specific (ds*FFluc*) or control (eGFP) dsRNA and luciferase activity determined. Reduced silencing was observed in cells expressing 3′UTR constructs compared to MBP-HDVr (Figure 5C). These results indicate that the 3′UTRs of LGTV and TBEV are able to interfere with the tick siRNA pathway.

DISCUSSION

RNAi is known to be a major defense mechanism against arboviruses in mosquitoes (10,11). Much less is known about ticks. Here, we investigated the antiviral RNAi response in *I. scapularis*-derived cells and viral counter-defense strategies. Our analysis reveals tick Ago and Dcr genes additional to those previously described (19). A significant gene expansion in the Ago subfamily has occurred in arachnids, compared to insects such as *D. melanogaster* and *A. aegypti*. Our results characterize key differences between *Ixodes* and mosquito RNAi responses. The antiviral activity of Ago-30, Ago-16 and Ago-68 (in case of vi-

ral infection) is in line with previous reports showing that mosquito/fly Ago-2 is involved in the antiviral RNAi response and phylogenetic analysis indicates that *Ixodes* Ago-30, Ago16 and Ago-68 are homologous to Ago-2 of insects (57,58). The expansion of putative Ago-2 paralogs in arachnids is different from other arthropods, which generally have one, or at most two, Ago-2 homologs. In contrast to Ago-16 and Ago-30, Ago-68 only shows antiviral activity in case of virus infection, which may suggests its involvement in limiting virus spread by pre-priming yet uninfected cells using systemic RNA silencing. Like mosquitoes, ticks appear to have undergone an expansion of the Piwi clade (Supplementary Figure S5C), though the expansion is smaller and occurred independently, in addition to a possible loss of Ago-3. We show that Dcr-90 is involved in antiviral RNAi against replicon in contrast to Dcr-89 showing significant antiviral activity in case of virus. However, failure of consistent/ efficient knockdown of Dcr-89 between experimental approaches and the borderline increase/significance of Dcr knockdowns on virus infection leaves it open whether or not a second Dcr protein is involved and if there are differences between effects of Dcr knockdowns on replicon and virus. Phylogenetic analysis, dependent on the rooting, maps Dcr-89 in a cluster with Dcr-2 proteins in insects. Dcr-2 is critical for the exogenous antiviral siRNA pathway in *Drosophila*, and presents a limiting factor for sufficient knockdown involving exogenous RNAi (using dsRNA) in this organism (14). Dcr-89 could act in a similar way in ticks, which may explain the lack of consistent knockdown. Dcr-90 showed an antiviral effect in IDE8 cells despite clustering with Dcr-1 proteins which have not yet been reported as antiviral in flies or mosquitoes. It cannot be excluded that potential antiviral functions of some Ago/Dcr proteins we describe here may have been missed due to inefficient knockdown of the transcript; however our results already show that mechanisms in ticks may differ in detail from those present in insects.

A key feature of antiviral RNAi in mosquitoes is the production of 21 nt viRNA molecules (10,11). The majority of viRNAs in IDE8 cells are 22 nt in length [as reported for viRNAs of the positive strand nodavirus in *C. elegans* (45,46)]. The same observation was made if an RNAi response was artificially induced by dsRNA. As the length of the siRNAs or viRNAs is mostly dependent on the Dcr enzyme, this indicates a key difference between *I. scapularis* and insect Dcr proteins. In insects, miRNA molecules differ from siRNA molecules (22 versus 21 nt) as they are mostly produced by Dcr-1. The antiviral effect of Dcr-90, which clusters with insect Dcr-1 proteins, and the production of 22 nt viRNAs points to differences between the antiviral RNAi pathways in *I. scapularis* and insects. Small RNAs of 22 nt were also found to be the major class of small RNA molecules that map to the genome of *I. scapularis*.

Little is known about the dsRNA substrate for Dcr-2 and the origin of viRNAs. Findings by us and others suggest that dsRNA replicative intermediates are Dcr-2 substrates in mosquitoes and derived cell lines and show that hot and cold spots of viRNAs are present along arbovirus genomes/antigenomes (21,38,40–43). This is in agreement with our results for SCRV and transfected dsRNA. In contrast, LGTV viRNAs map at highest frequencies to or around the 5′ and 3′ termini. In contrast, similar regions present in DENV and WNV are not particularly targeted by the RNAi machinery in mosquitoes (41–43). It has to be mentioned that recent work has shown that certain hot and cold spot observations are due to cloning bias of the small RNAs (59,60). The presence of small RNAs mapping to the non-coding strand of SCRV with a similar frequency as to the coding strand, supports the dsRNA genome as inducer molecule even with cloning bias. The same dsRNA-mediated induction may explain the bias of targeting the 5′ and 3′ genome termini of LGTV in IDE8 cells. A previously described replication-incompetent TBEV replicon (C17Fluc NS5 GAA) (23) behaved similar as the corresponding wild-type replicon in IDE8 cells with regards to luciferase production over time [in contrast to BHK where it shows reduction of luciferase production overtime as previously reported (23)] and production/ mapping of TBEV-specific small RNAs (Supplementary Figure S8 and Table 2). This suggests replication of the GAA mutant either by the viral replicase or other enzymes with complementing or replicative activity present in the IDE8 cells. Therefore such a mutant can unfortunately not be used to determine whether the observed TBEV-specific small RNAs are produced from incoming single stranded RNA, dsRNA replication intermediates or partial dsRNA.

Differences in the number of cells targeted by replicon or virus and the amount of virus/replicon RNA per cell could explain the difference in production of overall LGTV-specific small RNAs for replicon versus virus-infected IDE8 cells. Infection by full-length virus may also hide and limit the antiviral RNAi response in IDE8 cells more efficiently than replicon RNA which misses the coding sequences for structural proteins. Besides, the presence of structural proteins and nucleotide sequence (and thus changes in overall length of the viral RNA) may explain the observation that distribution of replicon viRNA versus virus shows some difference. Despite these differences though, LGTV viRNAs share common characteristics (bias for 22 nts viRNAs and targeting areas around the 5′ and 3′ genome termini) which are different to flavivirus-specific viRNAs reported in mosquitoes (41–43).

The detection of LGTV-specific viRNAs indicates the ability of the RNAi response to target the virus, raising the question: how can the virus still replicate in tick cells? Plant and 'true insect' viruses encode RSS proteins that interfere with the antiviral RNAi to allow successful viral infection (14,61). No arbovirus RSS protein is known, but an evasion strategy has been suggested for the alphavirus SFV (21) and the sfRNA molecules of mosquito-borne viruses interfere with RNAi responses (22). The 3′UTRs of tick- and mosquito-borne flaviviruses do not share high similarity at the nucleotide level and exchanging these sequences mostly leads to replication-deficient viruses (62–64). Despite this, bioinformatic modeling suggested a highly similar secondary RNA structure profile in the 3′ UTR of arthropod-borne flaviviruses, production and interference with the RNAi response was shown of TBEV and LGTV. WNV sfRNA is believed to mediate RSS activity by acting as a competitive substrate for Dcr (22). In contrast to WNV and DENV UTRs that do not appear to be specif-

ically targeted by Dcr (41–43), the 3′ UTR of the LGTV and TBEV replicon in IDE8 cells appears to be a target for Dcr activities; along with the 5′ UTR it generates the highest frequency of viRNAs. The sfRNA RSS activity probably results in less powerful activity than the known protein-based RSS of insect viruses. Expression of an RSS protein by alphaviruses results in reduced mosquito survival (40,65). Using a weak suppressor such as sfRNA may allow for sufficient levels of replication needed for successful transmission.

Taken together, our findings define details of the tick antiviral RNAi response and its interference by tick-borne arboviruses. They show several important differences in antiviral RNAi between different classes of arbovirus vectors (*Arachnida* versus *Insecta*) and broaden our knowledge about arthropod antiviral RNAi.

ACCESSION NUMBERS

ERP006219.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENT

The authors would like to thank Franz X. Heinz (Medical University of Vienna, Austria) for the pTNd/ΔME plasmid and TBEV replicons (C17Fluc and derived NS5 GAA mutant; C17 Fluc NS5 GAA), A. Pletnev (National Institutes of Health, USA) for the LGTV E5 cDNA clone and Ulrike Munderloh (University of Minnesota, USA) for the IDE8 cell line.

FUNDING

Netherlands Organisation for Scientific Research NWO [Rubicon fellowship, 825.10.021 to E.S.]; UK Biotechnology and Biological Sciences Research Council [Roslin Institute Strategic Programme Grant to J.K.F. and A.K.]; UK Medical Research Council [to A.K. and E.S.]; Wellcome Trust [Biomedical Resources Grant 088588 to J.K.F. and L.B.S., RCDF 085064/Z/08/Z to D.O.]; FP7-PEOPLE–ITN programme [EU Grant No. 238511 POSTICK ITN to S.W.]; Czech Science Foundation (GACR) [P302/12/2490 to H.T.], National Institutes of Health [AIO055672 to R.J.K.]; Division of Intramural Research, National Institutes of Health, National Institute of Allergy and Infectious Diseases [to S.M.B.]. Funding for open access charge: UK Medical Research Council.
Conflict of interest statement. None declared.

REFERENCES

1. Gritsun,T.S., Nuttall,P.A. and Gould,E.A. (2003) Tick-borne flaviviruses. *Adv. Virus Res.*, **61**, 317–371.
2. Best,S.M., Morris,K.L., Shannon,J.G., Robertson,S.J., Mitzel,D.N., Park,G.S., Boer,E., Wolfinbarger,J.B. and Bloom,M.E. (2005) Inhibition of interferon-stimulated JAK-STAT signaling by a tick-borne flavivirus and identification of NS5 as an interferon antagonist. *J. Virol.*, **79**, 12828–12839.
3. Park,G.S., Morris,K.L., Hallett,R.G., Bloom,M.E. and Best,S.M. (2007) Identification of residues critical for the interferon antagonist function of Langat virus NS5 reveals a role for the RNA-dependent RNA polymerase domain. *J. Virol.*, **81**, 6936–6946.
4. Pletnev,A.G. (2001) Infectious cDNA clone of attenuated Langat tick-borne flavivirus (strain E5) and a 3′ deletion mutant constructed from it exhibit decreased neuroinvasiveness in immunodeficient mice. *Virology*, **282**, 288–300.
5. Lindenbach,B.D. and Rice,C.M. (2003) Molecular biology of flaviviruses. *Adv. Virus Res.*, **59**, 23–61.
6. Villordo,S.M. and Gamarnik,A.V. (2009) Genome cyclization as strategy for flavivirus RNA replication. *Virus Res.*, **139**, 230–239.
7. Gritsun,T.S. and Gould,E.A. (2007) Origin and evolution of flavivirus 5′UTRs and panhandles: trans-terminal duplications? *Virology*, **366**, 8–15.
8. Gritsun,T.S. and Gould,E.A. (2007) Origin and evolution of 3′UTR of flaviviruses: long direct repeats as a basis for the formation of secondary structures and their significance for virus transmission. *Adv. Virus Res.*, **69**, 203–248.
9. Bell-Sakyi,L., Kohl,A., Bente,D.A. and Fazakerley,J.K. (2011) Tick cell lines for study of crimean-congo hemorrhagic fever virus and other arboviruses. Vector Borne Zoonotic Dis, **12**, 769–781.
10. Donald,C.L., Kohl,A. and Schnettler,E. (2012) New insights into control of arbovirus replication and spread by insect RNA interference pathways. *Insects*, **3**, 511–531.
11. Blair,C.D. (2011) Mosquito RNAi is the major innate immune pathway controlling arbovirus infection and transmission. *Future Microbiol.*, **6**, 265–277.
12. Fragkoudis,R., Attarzadeh-Yazdi,G., Nash,A.A., Fazakerley,J.K. and Kohl,A. (2009) Advances in dissecting mosquito innate immune responses to arbovirus infection. *J. Gen. Virol.*, **90**, 2061–2072.
13. Charrel,R.N., Attoui,H., Butenko,A.M., Clegg,J.C., Deubel,V., Frolova,T.V., Gould,E.A., Gritsun,T.S., Heinz,F.X., Labuda,M. *et al.* (2004) Tick-borne virus diseases of human interest in Europe. *Clin. Microbiol. Infect.*, **10**, 1040–1055.
14. Kemp,C. and Imler,J.L. (2009) Antiviral immunity in drosophila. *Curr. Opin. Immunol.*, **21**, 3–9.
15. de la Fuente,J., Kocan,K.M., Almazan,C. and Blouin,E.F. (2007) RNA interference for the study and genetic manipulation of ticks. *Trends Parasitol.*, **23**, 427–433.
16. Barry,G., Alberdi,P., Schnettler,E., Weisheit,S., Kohl,A., Fazakerley,J.K. and Bell-Sakyi,L. (2013) Gene silencing in tick cell lines using small interfering or long double-stranded RNA. *Exp. Appl. Acarol.*, **59**, 319–338.
17. Garcia,S., Billecocq,A., Crance,J.M., Munderloh,U., Garin,D. and Bouloy,M. (2005) Nairovirus RNA sequences expressed by a Semliki Forest virus replicon induce RNA interference in tick cells. *J. Virol.*, **79**, 8942–8947.
18. Garcia,S., Billecocq,A., Crance,J.M., Prins,M., Garin,D. and Bouloy,M. (2006) Viral suppressors of RNA interference impair RNA silencing induced by a Semliki Forest virus replicon in tick cells. *J. Gen. Virol.*, **87**, 1985–1989.
19. Kurscheid,S., Lew-Tabor,A.E., Rodriguez Valle,M., Bruyeres,A.G., Doogan,V.J., Munderloh,U.G., Guerrero,F.D., Barrero,R.A. and Bellgard,M.I. (2009) Evidence of a tick RNAi pathway by comparative genomics and reverse genetics screen of targets with known loss-of-function phenotypes in Drosophila. *BMC Mol. Biol.*, **10**, 26.
20. Ding,S.W. (2010) RNA-based antiviral immunity. *Nat. Rev. Immunol.*, **10**, 632–644.
21. Siu,R.W., Fragkoudis,R., Simmonds,P., Donald,C.L., Chase-Topping,M.E., Barry,G., Attarzadeh-Yazdi,G., Rodriguez-Andres,J., Nash,A.A., Merits,A. *et al.* (2011) Antiviral RNA interference responses induced by Semliki Forest virus infection of mosquito cells: characterization, origin, and frequency-dependent functions of virus-derived small interfering RNAs. *J. Virol.*, **85**, 2907–2917.
22. Schnettler,E., Sterken,M.G., Leung,J.Y., Metz,S.W., Geertsema,C., Goldbach,R.W., Vlak,J.M., Kohl,A., Khromykh,A.A. and Pijlman,G.P. (2012) Noncoding flavivirus RNA displays RNA interference suppressor activity in insect and Mammalian cells. *J. Virol.*, **86**, 13486–13500.
23. Hoenninger,V.M., Rouha,H., Orlinger,K.K., Miorin,L., Marcello,A., Kofler,R.M. and Mandl,C.W. (2008) Analysis of the effects of

alterations in the tick-borne encephalitis virus 3′-noncoding region on translation and RNA replication using reporter replicons. *Virology*, **377**, 419–430.
24. Gehrke,R., Ecker,M., Aberle,S.W., Allison,S.L., Heinz,F.X. and Mandl,C.W. (2003) Incorporation of tick-borne encephalitis virus replicons into virus-like particles by a packaging cell line. *J. Virol.*, **77**, 8924–8933.
25. Varela,M., Schnettler,E., Caporale,M., Murgia,C., Barry,G., McFarlane,M., McGregor,E., Piras,I.M., Shaw,A., Lamm,C. *et al.* (2013) Schmallenberg virus pathogenesis, tropism and interaction with the innate immune system of the host. *PLoS Pathog.*, **9**, e1003133.
26. Munderloh,U.G. and Kurtti,T.J. (1989) Formulation of medium for tick cell culture. *Exp. Appl. Acarol.*, **7**, 219–229.
27. Bell-Sakyi,L. (2004) Ehrlichia ruminantium grows in cell lines from four ixodid tick genera. *J. Comp. Pathol.*, **130**, 285–293.
28. Schnettler,E., Ratinier,M., Watson,M., Shaw,A.E., McFarlane,M., Varela,M., Elliott,R.M., Palmarini,M. and Kohl,A. (2013) RNA interference targets arbovirus replication in culicoides cells. *J. Virol.*, **87**, 2441–2454.
29. Watson,M., Schnettler,E. and Kohl,A. (2013) viRome: an R package for the visualization and analysis of viral small RNA sequence datasets. *Bioinformatics*, **29**, 1902–1903.
30. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
31. Loytynoja,A. and Goldman,N. (2010) webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, **11**, 579.
32. Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.
33. Ronquist,F., Teslenko,M., van der Mark,P., Ayres,D.L., Darling,A., Hohna,S., Larget,B., Liu,L., Suchard,M.A. and Huelsenbeck,J.P. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, **61**, 539–542.
34. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
35. Hemmes,H., Kaaij,L., Lohuis,D., Prins,M., Goldbach,R. and Schnettler,E. (2009) Binding of small interfering RNA molecules is crucial for RNA interference suppressor activity of rice hoja blanca virus NS3 in plants. *J. Gen. Virol.*, **90**, 1762–1766.
36. Campbell,C.L., Keene,K.M., Brackney,D.E., Olson,K.E., Blair,C.D., Wilusz,J. and Foy,B.D. (2008) Aedes aegypti uses RNA interference in defense against Sindbis virus infection. *BMC Microbiol.*, **8**, 47.
37. Flynt,A., Liu,N., Martin,R. and Lai,E.C. (2009) Dicing of viral replication intermediates during silencing of latent Drosophila viruses. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 5270–5275.
38. Morazzani,E.M., Wiley,M.R., Murreddu,M.G., Adelman,Z.N. and Myles,K.M. (2012) Production of virus-derived ping-pong-dependent piRNA-like small RNAs in the mosquito soma. *PLoS Pathog.*, **8**, e1002470.
39. Myles,K.M., Morazzani,E.M. and Adelman,Z.N. (2009) Origins of alphavirus-derived small RNAs in mosquitoes. *RNA Biol.*, **6**, 387–391.
40. Myles,K.M., Wiley,M.R., Morazzani,E.M. and Adelman,Z.N. (2008) Alphavirus-derived small RNAs modulate pathogenesis in disease vector mosquitoes. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 19938–19943.
41. Scott,J.C., Brackney,D.E., Campbell,C.L., Bondu-Hawkins,V., Hjelle,B., Ebel,G.D., Olson,K.E. and Blair,C.D. (2010) Comparison of dengue virus type 2-specific small RNAs from RNA interference-competent and -incompetent mosquito cells. *PLoS Negl. Trop. Dis.*, **4**, e848.
42. Brackney,D.E., Beane,J.E. and Ebel,G.D. (2009) RNAi targeting of West Nile virus in mosquito midguts promotes virus diversification. *PLoS Pathog.*, **5**, e1000502.
43. Brackney,D.E., Scott,J.C., Sagawa,F., Woodward,J.E., Miller,N.A., Schilkey,F.D., Mudge,J., Wilusz,J., Olson,K.E., Blair,C.D. *et al.* (2010) C6/36 Aedes albopictus cells have a dysfunctional antiviral RNA interference response. *PLoS. Negl. Trop. Dis.*, **4**, e856.
44. Ding,S.W. and Lu,R. (2011) Virus-derived siRNAs and piRNAs in immunity and pathogenesis. *Curr. Opin. Virol.*, **1**, 533–544.
45. Felix,M.A., Ashe,A., Piffaretti,J., Wu,G., Nuez,I., Belicard,T., Jiang,Y., Zhao,G., Franz,C.J., Goldstein,L.D. *et al.* (2011) Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses. *PLoS Biol.*, **9**, e1000586.
46. Wu,Q., Luo,Y., Lu,R., Lau,N., Lai,E.C., Li,W.X. and Ding,S.W. (2010) Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc. Natl. Acad. Sci. U.S.A*, **107**, 1606–1611.
47. Aliyari,R., Wu,Q., Li,H.W., Wang,X.H., Li,F., Green,L.D., Han,C.S., Li,W.X. and Ding,S.W. (2008) Mechanism of induction and suppression of antiviral immunity directed by virus-derived small RNAs in Drosophila. *Cell Host Microbe*, **4**, 387–397.
48. Alberdi,M.P., Dalby,M.J., Rodriguez-Andres,J., Fazakerley,J.K., Kohl,A. and Bell-Sakyi,L. (2012) Detection and identification of putative bacterial endosymbionts and endogenous viruses in tick cell lines. *Ticks Tick Borne Dis.*, **3**, 137–146.
49. Attoui,H., Stirling,J.M., Munderloh,U.G., Billoir,F., Brookes,S.M., Burroughs,J.N., de Micco,P., Mertens,P.P. and de Lamballerie,X. (2001) Complete sequence characterization of the genome of the St Croix River virus, a new orbivirus isolated from cells of Ixodes scapularis. *J. Gen. Virol.*, **82**, 795–804.
50. Aliyari,R. and Ding,S.W. (2009) RNA-based viral immunity initiated by the Dicer family of host immune receptors. *Immunol. Rev.*, **227**, 176–188.
51. Cerutti,H. and Casas-Mollano,J.A. (2006) On the origin and functions of RNA-mediated silencing: from protists to man. *Curr. Genet.*, **50**, 81–99.
52. Funk,A., Truong,K., Nagasaki,T., Torres,S., Floden,N., Balmori Melian,E., Edmonds,J., Dong,H., Shi,P.Y. and Khromykh,A.A. (2010) RNA structures required for production of subgenomic flavivirus RNA. *J. Virol.*, **84**, 11407–11417.
53. Pijlman,G.P., Funk,A., Kondratieva,N., Leung,J., Torres,S., van der Aa,L., Liu,W.J., Palmenberg,A.C., Shi,P.Y., Hall,R.A. *et al.* (2008) A highly structured, nuclease-resistant, noncoding RNA produced by flaviviruses is required for pathogenicity. *Cell Host Microbe*, **4**, 579–591.
54. Silva,P.A., Pereira,C.F., Dalebout,T.J., Spaan,W.J. and Bredenbeek,P.J. (2010) An RNA pseudoknot is required for production of yellow fever virus subgenomic RNA by the host nuclease XRN1. *J. Virol.*, **84**, 11395–11406.
55. Markoff,L. (2003) 5′- and 3′-noncoding regions in flavivirus RNA. *Adv. Virus Res.*, **59**, 177–228.
56. Mitzel,D.N., Best,S.M., Masnick,M.F., Porcella,S.F., Wolfinbarger,J.B. and Bloom,M.E. (2008) Identification of genetic determinants of a tick-borne flavivirus associated with host-specific adaptation and pathogenicity. *Virology*, **381**, 268–276.
57. Keene,K.M., Foy,B.D., Sanchez-Vargas,I., Beaty,B.J., Blair,C.D. and Olson,K.E. (2004) RNA interference acts as a natural antiviral response to O'nyong-nyong virus (Alphavirus; Togaviridae) infection of Anopheles gambiae. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 17240–17245.
58. van Rij,R.P., Saleh,M.C., Berry,B., Foo,C., Houk,A., Antoniewski,C. and Andino,R. (2006) The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster. *Genes Dev.*, **20**, 2985–2995.
59. Sorefan,K., Pais,H., Hall,A.E., Kozomara,A., Griffiths-Jones,S., Moulton,V. and Dalmay,T. (2012) Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence*, **3**, 4.
60. Zhuang,F., Fuchs,R.T., Sun,Z., Zheng,Y. and Robb,G.B. (2012) Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. *Nucleic Acids Res.*, **40**, e54.
61. Li,F. and Ding,S.W. (2006) Virus counterdefense: diverse strategies for evading the RNA-silencing immunity. *Annu. Rev. Microbiol.*, **60**, 503–531.
62. Friebe,P., Shi,P.Y. and Harris,E. (2011) The 5′ and 3′ downstream AUG region elements are required for mosquito-borne flavivirus RNA replication. *J. Virol.*, **85**, 1900–1905.
63. Romero,T.A., Tumban,E., Jun,J., Lott,W.B. and Hanley,K.A. (2006) Secondary structure of dengue virus type 4 3′ untranslated region: impact of deletion and substitution mutations. *J. Gen. Virol.*, **87**, 3291–3296.
64. Tumban,E., Mitzel,D.N., Maes,N.E., Hanson,C.T., Whitehead,S.S. and Hanley,K.A. (2011) Replacement of the 3′ untranslated variable

region of mosquito-borne dengue virus with that of tick-borne Langat virus does not alter vector specificity. *J. Gen. Virol.*, **92**, 841–848.
65. Cirimotich,C.M., Scott,J.C., Phillips,A.T., Geiss,B.J. and Olson,K.E. (2009) Suppression of RNA interference increases alphavirus replication and virus-associated mortality in Aedes aegypti mosquitoes. *BMC Microbiol.*, **9**, 49.

Journal of General Virology (2013), **94**, 1680–1689 **DOI** 10.1099/vir.0.053850-0

Knockdown of piRNA pathway proteins results in enhanced Semliki Forest virus production in mosquito cells

Esther Schnettler,[1,2] Claire L. Donald,[1,2] Stacey Human,[1,3] Mick Watson,[4] Ricky W. C. Siu,[1]† Melanie McFarlane,[2] John K. Fazakerley,[1,3] Alain Kohl[1,2] and Rennos Fragkoudis[1,3]

Correspondence
Esther Schnettler
Esther.Schnettler@glasgow.ac.uk
Rennos Fragkoudis
rennos.fragkoudis@pirbright.ac.uk

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK

[2]MRC-University of Glasgow Centre for Virus Research, 8 Church Street, Glasgow G11 5JR, UK

[3]The Pirbright Institute, Ash Road, Pirbright, Surrey GU24 0NF, UK

[4]ARK Genomics, The Roslin Institute, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK

The exogenous siRNA pathway is important in restricting arbovirus infection in mosquitoes. Less is known about the role of the PIWI-interacting RNA pathway, or piRNA pathway, in antiviral responses. Viral piRNA-like molecules have recently been described following infection of mosquitoes and derived cell lines with several arboviruses. The piRNA pathway has thus been suggested to function as an additional small RNA-mediated antiviral response to the known infection-induced siRNA response. Here we show that piRNA-like molecules are produced following infection with the naturally mosquito-borne Semliki Forest virus in mosquito cell lines. We show that knockdown of piRNA pathway proteins enhances the replication of this arbovirus and defines the contribution of piRNA pathway effectors, thus characterizing the antiviral properties of the piRNA pathway. In conclusion, arbovirus infection can trigger the piRNA pathway in mosquito cells, and knockdown of piRNA proteins enhances virus production.

Received 26 March 2013
Accepted 27 March 2013

INTRODUCTION

Arboviruses are unique in that they must naturally replicate in both their invertebrate vector and vertebrate host and are therefore subjected to the selective pressure of very different antiviral responses. One of the major antiviral responses in invertebrates is the RNA silencing pathway or RNA interference (RNAi). It has been shown that the RNAi pathway, in particular the exogenous small interfering (si)RNA pathway, is able to inhibit and restrict arbovirus infections in whole mosquitoes or mosquito cells (Blair, 2011; Donald *et al.*, 2012). The exogenous RNAi pathway is induced by virus-derived dsRNA that is recognized by a Dicer protein, Dcr-2, and is processed into 21 bp-long virus-derived siRNAs, also called viRNAs. After viRNAs are incorporated and unwound in the RNA-induced silencing complex (RISC) that harbours Argonaute 2 (Ago 2) as a catalytic domain, one strand of the viRNA is retained and used as a guide to find complementary viral RNA, which is then degraded. Until recently, it was believed that the antiviral response in invertebrates is only attributed to the exogenous siRNA pathway. Recently, however, the PIWI-interacting RNA (piRNA) pathway has also been suggested to display antiviral activity. piRNA molecules differ from siRNAs in several aspects; they are produced by a Dicer-independent pathway; have a broader size range of 25–29 nt; are associated with proteins of the PIWI clade and have a so-called 'ping-pong' signature due to their production pathway, which is represented by a bias for U at position 1 in antisense piRNAs and A at position 10 in sense piRNAs (Saito & Siomi, 2010; Senti & Brennecke, 2010; Siomi *et al.*, 2010, 2011; van Rij & Berezikov, 2009). In *Drosophila melanogaster*, it has been shown that PIWI proteins are mainly expressed in germline cells and are thought to protect the germline from transposable elements by targeting the transcribed RNA of active transposons. However, PIWI proteins have also been detected in somatic cells (Brennecke *et al.*, 2007). Although their induction pathway is still not completely understood, two mechanisms have been proposed to describe piRNA biogenesis. Primary piRNA molecules are

†Present address: Dalhousie University, Department of Microbiology & Immunology/Paediatrics, 5850 College Street, Halifax B3H 1X5, Canada.

Two supplementary figures and one table are available with the online version of this paper.

antisense to the genomic regions of transposons and derive from long precursor ssRNA that targets transposon-derived sense RNA. Upon cleavage, they give rise to secondary piRNA molecules that are mostly sense with an A_{10} bias. Secondary piRNAs are incorporated into Argonaute 3 (Ago 3) protein, which uses these piRNAs to find complementary antisense RNA, which again results in the production of primary-type piRNAs. This so-called ping-pong mechanism results in the generation of anti-sense primary piRNA molecules with a U_1 bias. Primary piRNA molecules have mostly been found to form complexes with Aubergine (Aub) and PIWI proteins (Saito & Siomi, 2010; Senti & Brennecke, 2010; Siomi *et al.*, 2010, 2011; van Rij & Berezikov, 2009).

The detection of virus-specific piRNA molecules in drosophila ovary somatic sheet (OSS) cells was the first report suggesting that the piRNA pathway targeted viruses in insects (Wu *et al.*, 2010). More recently, virus-specific piRNA molecules have been reported in aedine mosquitoes for chikungunya virus (CHIKV) (*Togaviridae*, *Alphavirus*) (*Aedes albopictus* and *Ae. aegypti*) and dengue virus (DENV) (*Flaviviridae*, *Flavivirus*) (*Ae. aegypti*), and their derived cell lines can become infected with Sindbis virus (SINV) (*Togaviridae*, *Alphavirus*), La Crosse virus (LACV) (*Bunyaviridae*, *Orthobunyavirus*), Rift Valley fever virus (RVFV) (*Bunyaviridae*, *Phlebovirus*) and Schmallenberg virus (SBV) (*Bunyaviridae*, *Orthobunyavirus*) (Hess *et al.*, 2011; Léger *et al.*, 2013; Morazzani *et al.*, 2012; Schnettler *et al.*, 2013; Vodovar *et al.*, 2012). It is not known whether these virus-specific piRNA molecules actually mediate any antiviral activities or which proteins of the piRNA pathway are important for this response. The PIWI protein clade shows an expansion in aedine mosquitoes compared to drosophila, which is consistent with a role besides transposon targeting. *Ae. aegypti* encode seven Piwi proteins (Piwi 1, AAEL008076; Piwi 2, AAEL008098; Piwi 3, AAEL013692; Piwi 4, AAEL007698; Piwi 5, AAEL013233; Piwi 6, AAEL013227; Piwi 7, AAEL006287) and one Ago 3 protein (AAEL007823), compared to *D. melanogaster*, which only encodes one of each of Piwi, Ago 3 and Aub (Campbell *et al.*, 2008a).

Although expression of some of the PIWI proteins has been recently reported in *Ae. aegypti*-derived Aag2 cell lines (Vodovar *et al.*, 2012) and in the head and thorax of *Ae. albopictus* (Morazzani *et al.*, 2012), nothing is known about their involvement in antiviral activity. If the piRNA pathway acts as an antiviral response, then it would be expected that silencing proteins involved would have a positive effect on arbovirus replication as observed for the Ago 2 protein, which is known to be involved in the siRNA-based antiviral RNAi response (Campbell *et al.*, 2008b; Sánchez-Vargas *et al.*, 2009). To test this hypothesis, we investigated the importance of piRNA-related proteins on viral infection. Re-analysis of previous deep-sequencing data from mosquito-borne Semliki Forest virus (SFV) (*Togaviridae*, *Alphavirus*) infection of U4.4 (derived from *Ae. albopictus*) or Aag2 (derived from *Ae. aegypti*) cells (Siu *et al.*, 2011) revealed the presence of piRNA-like small RNAs mapping mainly to a section of the SFV genome, which decreased in Aag2 cells following knockdown for all Piwi/Ago 3 proteins. Silencing of PIWI 4 protein increased SFV replication and production but did not decrease the presence of SFV-specific piRNA-like molecules, confirming that the piRNA pathway does indeed display antiviral activity and that Piwi 4 possibly acts as an antiviral effector protein in this pathway.

RESULTS

SFV-specific piRNA-like molecules in aedine cell lines

To determine whether the piRNA pathway specifically targets SFV in mosquito cells, we first investigated if these cells produce viral-specific piRNA-like molecules following infection. We re-analysed data previously obtained from deep sequencing of Aag2 and U4.4 cells infected with SFV [RNA isolation 24 h post-infection (p.i.)]; deep sequencing by using the Illumina Solexa platform as described in Methods (Siu *et al.*, 2011) and this time also mapped small RNAs greater than 26 nt in length to the SFV genome. As previously reported, the major species of virus-specific small RNA molecules were viRNAs 21 nt in length (Siu *et al.*, 2011); however, small RNAs mapping to SFV in the range of 25–29 nt could be observed for both cell lines (Fig. 1a). Most of these small RNA molecules mapped to the sense orientation of the SFV in the 5′ end of the subgenomic RNA and had a bias for A at position 10, a characteristic of secondary piRNAs (Fig. 1b and c). Besides, the 5′ ends of these complementary SFV-specific RNAs were most frequently separated by 10 nt, a feature of piRNAs produced by the ping-pong mechanism (Fig. 1d) This is consistent with what has previously been reported for SINV, CHIKV and SBV-specific piRNA-like RNAs (Morazzani *et al.*, 2012; Schnettler *et al.*, 2013; Vodovar *et al.*, 2012). Having demonstrated the production of piRNA-like RNAs in our mosquito cell infection systems, we proceeded to investigate piRNA pathway functionality by determining the effect of Piwi/Ago 3 silencing on SFV replication.

Expression and knockdown of PIWI transcripts in Aag2 cells

Given the lack of genomic information for *Ae. albopictus*, these experiments were performed in *Ae. aegypti*-derived Aag2 cells. To produce dsRNA molecules specifically targeting single PIWIs or Ago 3, primers were designed to amplify unique regions of these genes by RT-PCR. Ago 2 depletion was taken as a positive control as it has been previously reported to be involved in the antiviral siRNA pathway (Campbell *et al.*, 2008b; Sánchez-Vargas *et al.*, 2009; van Rij *et al.*, 2006), and Ago 1 was a negative control that is known to be involved in the microRNA pathway.

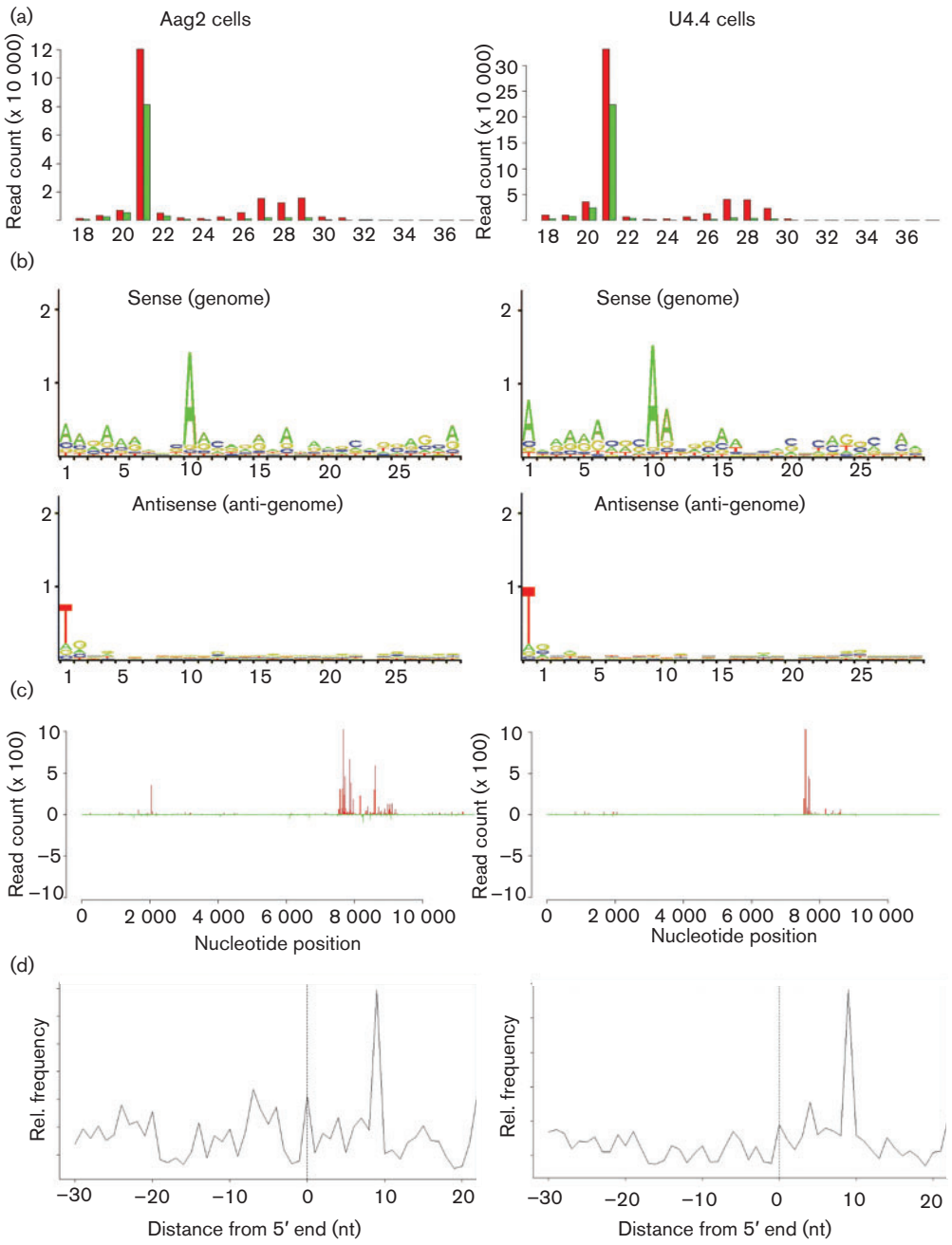


Fig. 1. Aag2 and U4.4 cells produce both viRNAs and piRNA-like RNAs following SFV infection. (a) Size distribution of small RNA molecules mapping to the SFV genome or anti-genome in *Ae. aegypti* (Aag2) or *Ae. albopictus* (U4.4); RNA was isolated at 24 h p.i. Red and green indicate small RNAs mapping to the genome and anti-genome, respectively. (b) Relative nt frequency and conservation per position of 25–29 nt small RNAs mapping to the genome and anti-genome of SFV in Aag2 and U4.4 cells are indicated. Sequence is represented as DNA. The overall height of the nt represents sequence conservation. (c) Frequency distribution of 28 nt small RNA molecules to the SFV genome or anti-genome in Aag2 and U4.4. The *y*-axis shows the frequency of the 28 nt small RNAs mapping to the corresponding nt position of the *x*-axis (SFV genome length). Positive numbers represent the frequency of small RNAs mapping to the genome and negative numbers those mapping to the anti-genome. (d) Frequency map of 24–30 nt small RNAs mapping to the opposite strand of SFV4. Probabilities of complementarities of the sense and antisense SFV-specific small RNAs were mapped along the small RNAs (position 0 represents the first nt).

First, the primers were tested for their specificity to amplify unique regions of the Piwi/Ago 3 mRNAs. As previously reported (Vodovar *et al.*, 2012), we amplified Piwi 4, 5, 6 and 7, as well as Ago 3. Piwi 1–3 are highly homologous, making unique primer design difficult. Primers amplifying parts shared by either Piwi 1, 2 and 3 or only 2 and 3 were successful, as well as Piwi 2 and 3 alone; however, attempts to amplify a unique region of Piwi 1 were unsuccessful (Fig. 2a). Sequencing of the PCR products confirmed their origin. Before, silencing the Piwi and Ago 3 with the dsRNA produced by *in vitro* transcription, transfection efficiency of dsRNA in Aag2 was assessed and optimized using internally labelled fluorescent dsRNA molecules. A maximum of 28.6 % positive cells was observed (Fig. S1a, available in JGV Online). Cells were transfected with 100 ng dsRNA, either Piwi specific (1/2/3, 2/3, 2, 3, 4, 5, 6, 7 and Ago 3) or control (eGFP specific), at 24 h post-seeding using Lipofectamine 2000. Silencing of target transcripts was determined by semi-quantitative reverse transcriptase PCR (RT-PCR) 24 h post-transfection (p.t.), and several experiments were quantified in relation to control dsRNA using actin as a loading control (Fig. 2c). Aag2 cells treated with dsRNA specific for Piwi 1/2/3, 2/3, 4, 5, 6 and Ago 3 showed a 10–42 % reduction in target transcripts compared to controls treated with eGFP dsRNA. Similar results were observed for Piwi 2, 3 and 7 (Fig. 2b, c). A cell viability assay (cellTiter-Glo, Promega) was performed on all dsRNA-treated cells to determine whether transcript knockdown had an effect on cell

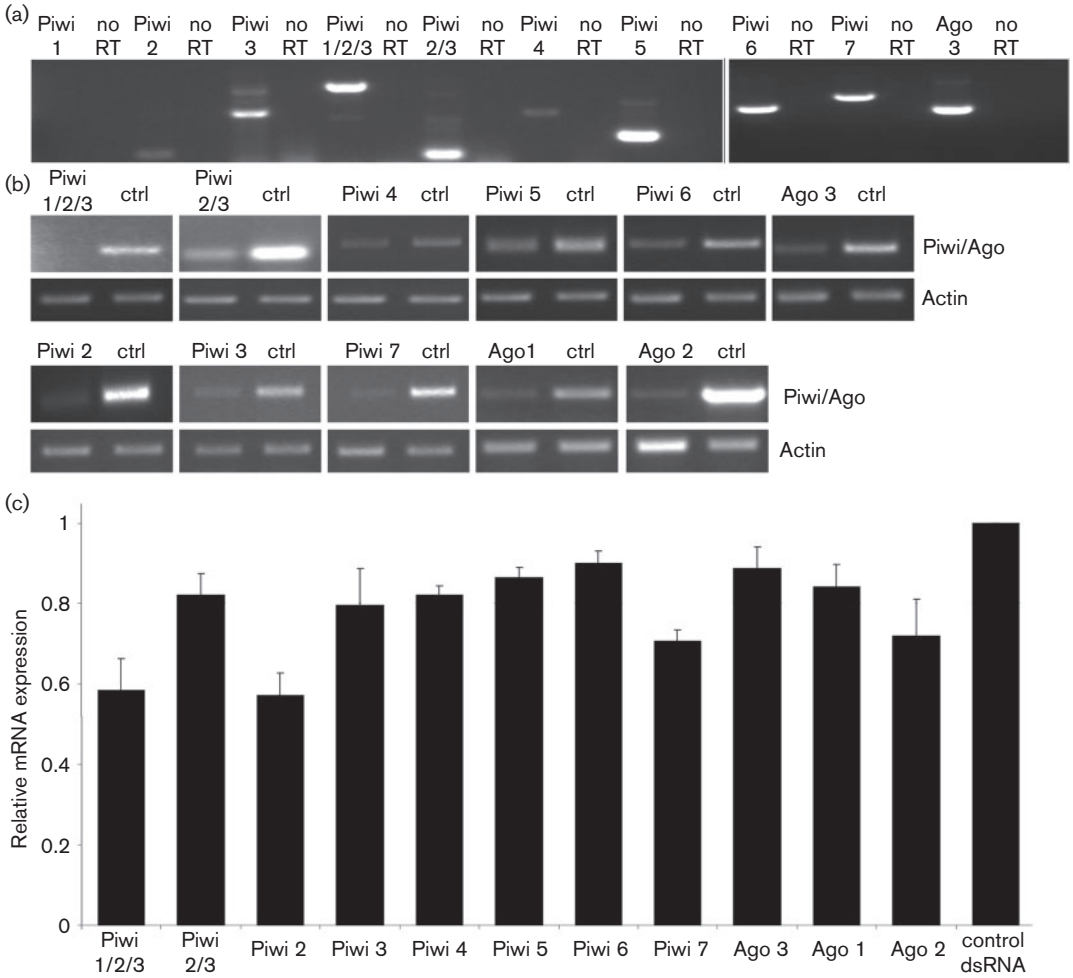


Fig. 2. Expression and knockdown of piRNA-related transcripts in Aag2 cells. (a) Detection of Piwi (1/2/3, 2/3, 2, 3, 4, 5, 6, 7) and Ago 3 transcripts in *Ae. aegypti*-derived Aag2 cells by RT-PCR using oligo-dT primers for reverse transcription, and gene-specific primers for PCR. no RT represents the PCR product derived from samples lacking the superscript III enzyme. (b) dsRNA-based silencing of Piwi (1/2/3, 2/3, 2, 3, 4, 5, 6 and 7), Ago 3, Ago 1 and Ago 2 transcripts or cells transfected with eGFP-specific control dsRNA (ctrl) were detected in Aag2 cells by RT-PCR using gene-specific primers. Actin PCR product was used as a control. (c) Quantification of mRNA knockdowns using ImageJ software (National Institutes of Health). Graph shows the mean expression of five repeats normalized to actin expression and relative to eGFP-dsRNA controls. Error bars show standard errors of means.

viability, but no deleterious effect was observed (data not shown).

Effect of Piwi/Ago 3 knockdown on SFV replication

Next, the effect of Piwi/Ago 3 silencing on SFV replication was investigated and compared to the knockdowns of Ago 1 and 2. dsRNA transfections in Aag2 cells were repeated, and at 24 h p.t., these cells were infected with the reporter alphavirus SFV4(3H)-*RLuc* [expressing *Renilla* luciferase (*RLuc*) as a replication marker] (Fig. 3a). Infections were performed at an m.o.i. of 0.1 (Fig. 3b, c), and *RLuc* activity was determined 48 h p.i. Significantly higher luciferase activity was detected in cells treated with Piwi 4-specific dsRNA compared to control (Fig. 3b, c). Cells treated with Piwi 6-, Piwi

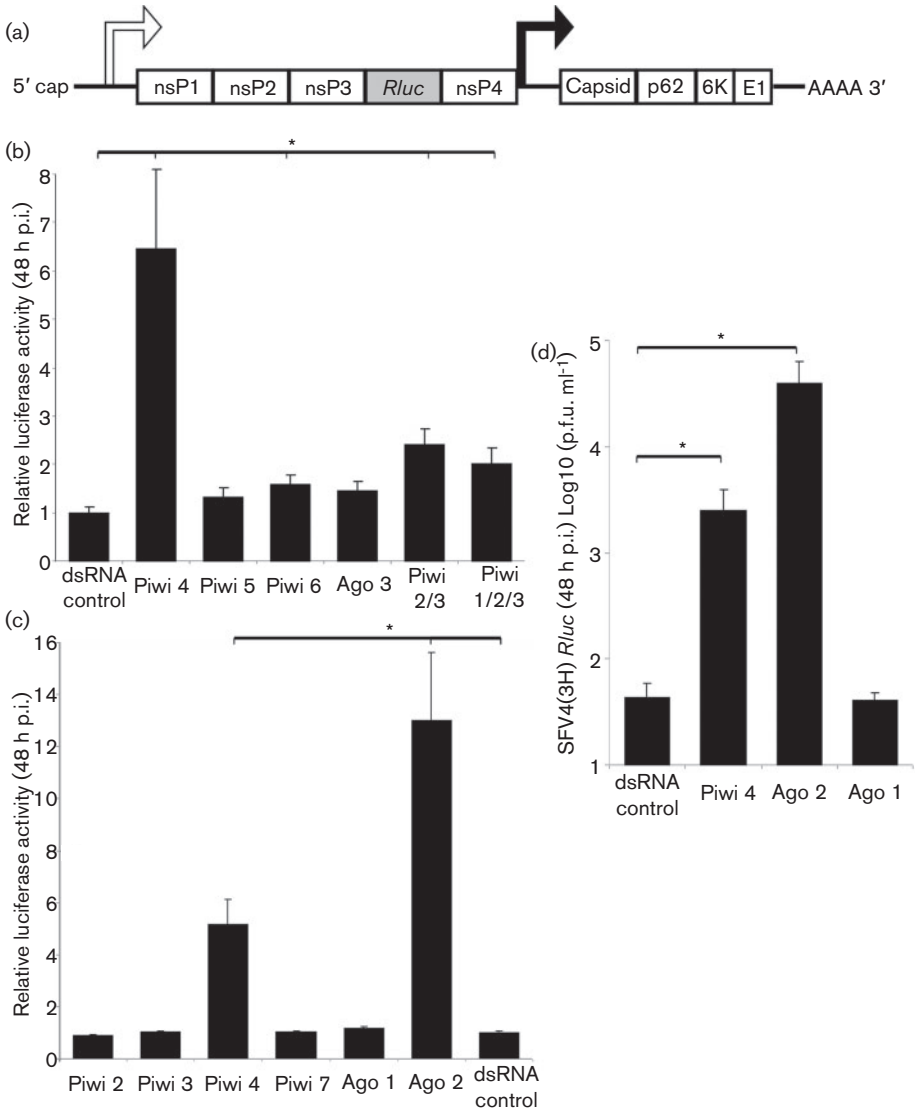


Fig. 3. Piwi/Ago 3 proteins inhibit SFV replication in Aag2 cells. (a) Schematic representation of SFV4 encoding *Renilla* luciferase (*RLuc*) as reporter (flanked by duplicated nsP2-protease cleavage sites at the nsP3/4 junction) as part of the viral non-structural polyprotein; SFV4(3H)-*RLuc* virus. (b) Aag2 cells transfected with dsRNA against Piwi (1/2/3, 2/3, 4, 5 and 6), Ago 3 or eGFP-specific dsRNA (control) were infected with SFV4(3H)-*RLuc* 24 h p.t. at an m.o.i. of 0.1. The mean of four independent experiments performed in triplicate are shown with standard errors (* represents $p<0.05$, *t*-test). (c) As (b) with dsRNA specific against Piwi (2, 3, 4 and 7), Ago 2, Ago 1 or eGFP-specific dsRNA (control). Luciferase activity was measured 48 h p.i., and the means with standard errors are shown for three independent experiments performed in triplicate (* represents $p<0.05$, *t*-test). (d) SFV titre (p.f.u. ml^{-1}) in supernatant of Piwi 4-, Ago 1- or Ago 2-silenced cells versus control (eGFP dsRNA) infected with an m.o.i. of 0.1 was determined 48 hp.i. by plaque assay. The means of three independent experiments performed in triplicate are shown with standard errors (* represents $p<0.05$, *t*-test).

2/3- and Piwi 1/2/3-specific dsRNA showed an increase in *RLuc* activity, although Piwi 4-specific dsRNA had a stronger effect (Fig. 3b). Knockdown of Ago 1 had no effect on luciferase expression compared to Ago 2 knockdowns, which exhibited the highest increase in luciferase activity (Fig. 3c). In addition, plaque assays performed with supernatant from dsRNA-transfected (eGFP, Piwi 4, Ago 1 and Ago 2) and SFV4(3H)-*RLuc*-infected cells (m.o.i. of 0.1) showed higher virus titre in cells treated with Piwi 4 or Ago 2 dsRNA (Fig. 3d). To ensure that the observed increase of *RLuc* activity in cells transfected with Piwi 4-specific dsRNA was not due to off-target effects of the dsRNA, experiments were repeated with two additional Piwi 4-specific dsRNA molecules (Piwi 4-2 and Piwi 4-3), resulting in similar *RLuc* activity (Fig. S1b). Overall, these results show that silencing Ago 2 and some Piwi, in particular Piwi 4, in Aag2 cells enhances SFV replication and virion production.

Effect of PIWI/Ago 3 knockdown on the production of SFV-specific piRNA-like molecules

Deep-sequencing experiments were performed to determine in more detail if Piwi 4 is needed for the production of SFV-specific piRNA-like molecules or rather acts as an effector molecule using the produced SFV-specific piRNA-like molecules to target the viral RNAs. Knockdown of all Piwi/Ago 3 proteins was performed to determined that any of these proteins are needed for the production of SFV-specific piRNA-like molecules. First, we established that the same cells could be targeted by dsRNA transfection and SFV infection using internally labelled fluorescent dsRNA and immunostaining for SFV nsP3 (Fig. S1a). Next, cells were transfected either with a combination of dsRNA molecules (targeting Piwi 1-3, 4, 5, 6, 7 and Ago 3) or Piwi 4-specific dsRNA alone, followed by SFV4 infection at an m.o.i. of 10. Cells transfected with eGFP-specific dsRNA were used as control. At 24 h p.i., total RNA was isolated, small RNAs were sequenced and the frequencies and SFV genome location of small RNAs were determined. All samples showed the presence of 21 nt SFV-specific small RNAs with a similar frequency to the genome and anti-genome; however, their frequency differs depending on the transfected dsRNAs, giving the highest number in cells transfected with a combination of piRNA/Ago 3-specific dsRNA, followed by Piwi 4-specific dsRNA, with control eGFP-specific dsRNA giving the lowest frequency. In addition, SFV-specific small RNAs of length 26–30 nt with a peak at 27 nt, mapping mainly to the sense orientation, could be observed in cells transfected with eGFP-specific control dsRNA and Piwi 4-specific dsRNA. Similar molecules were also present in cells transfected with a combination of Piwi/Ago 3-specific dsRNA but at a much lower frequency. These molecules have all the piRNA-specific features described for the previously identified SFV-specific small RNA molecules: characterized by an A_{10} bias in the sense molecules, a U_1 bias in the antisense molecules (Fig. 4a, b) and separation of 10 nt of the 5′ends of the complementary small RNAs (Fig. S2b). As already observed for the other SFV-specific piRNA-like molecules, they mainly map to the 5′ end of the subgenomic RNA (Fig. 2a). To further characterize the response of SFV replication in these knockdown cells, the experiments were repeated following infection with the SFV4(3H)-*RLuc* reporter virus. Cells transfected with a combination of Piwi/Ago 3-specific dsRNA molecules, lacking Piwi 4 dsRNA, were also included. Increase in *RLuc* activity compared to control cells could be observed for all knockdowns; however, the strongest increase was present in cells with Piwi 4 knockdown followed by knockdown of all Piwi/Ago 3. Interestingly, cells transfected with a combination of Piwi/Ago 3-specific dsRNA but lacking Piwi 4-specific dsRNAs resulted in the lowest *RLuc* increase (Fig. 4c). Overall, these results support the involvement of Piwi/Ago 3 for the production of the SFV-specific piRNA-like molecules and suggest that Piwi 4 acts as an effector protein that targets the virus but is not needed for the production of SFV-specific piRNA-like molecules.

Can dsRNA molecules induce piRNA production

Knockdown experiments of Ago 3 performed in *Anopheles gambiae* suggests that at least some of the PIWI pathway proteins are also involved in the exogenous dsRNA-induced silencing response (Hoa *et al.*, 2003). In addition, recent experiments in aedine mosquitoes infected with transgenic CHIKV expressing the dsRNA-binding protein B2 suggest that dsRNA molecules are an inducer of the piRNA pathway (Morazzani *et al.*, 2012), as is known for the exogenous siRNA pathway. To investigate if dsRNA on its own can be processed into piRNA-like molecules, we transfected Aag2 or U4.4 cells with dsRNA molecules derived from the eGFP sequence. Subsequently, RNA was isolated 24 h p.t., followed by sequencing and mapping of small RNAs to the eGFP target sequence as described above. As expected, small RNAs of 21 nt in size that mapped to the sense or antisense orientation and along the eGFP sequence were observed as the majority, indicating induction of the exogenous RNAi pathway (Fig. 5a, b). Some small RNAs in the 25–29 nt range mapping to the eGFP input sequence were identified; however, no specific sequence logo identifying them as piRNA-like molecules was detected (data not shown). This suggests that ssRNA (for example from virus replication) is needed for the production of piRNA molecules but does not rule out a link between the siRNA and piRNA pathways.

DISCUSSION

Until now, antiviral RNA silencing activities in mosquitoes have mainly been reported for the exogenous siRNA pathway. The identification of piRNA-like virus-specific RNA molecules in drosophila OSS (Wu *et al.*, 2010), mosquitoes and mosquito-derived cells against different arboviruses suggested a contribution of the piRNA pathway in the antiviral response (Hess *et al.*, 2011; Léger

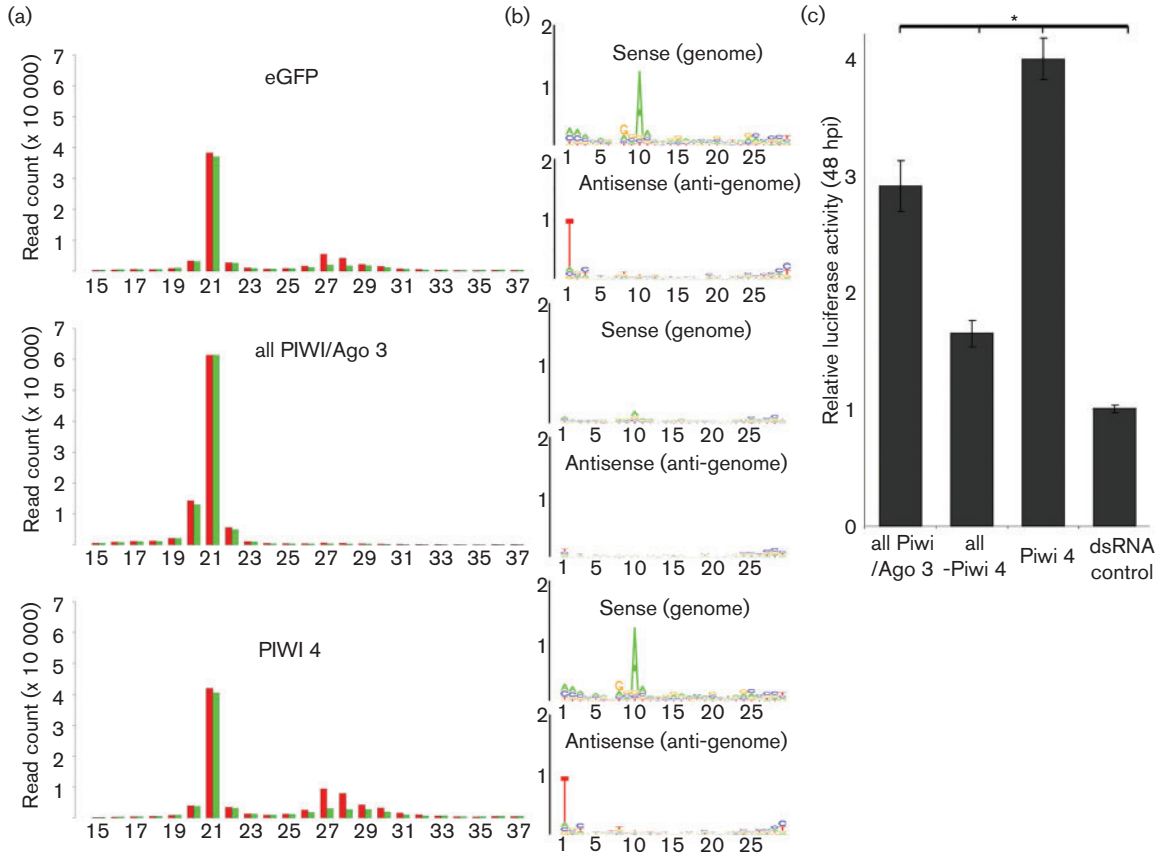


Fig. 4. Piwi/Ago 3 proteins are involved in the production of SFV-specific piRNA-like molecules in Aag2 cells. (a) Size distribution of small RNA molecules mapping to the SFV genome or anti-genome in Aag2 transfected with eGFP-specific control dsRNA, a combination of Piwi/Ago 3 dsRNA (Piwi 1/2/3, 2/3, 2, 3, 4, 5, 6, 7 and Ago 3) or Piwi 4-specific dsRNA, followed by SFV4 infection; RNA was isolated at 24 h.p.i. Red and green indicate small RNAs mapping to the genome and anti-genome, respectively. (b) Relative nt frequency and conservation per position of 25–29 nt small RNAs mapping to the genome and anti-genome of SFV in Aag2 prior to transfection with the above-mentioned dsRNA molecules. Sequence is represented as DNA. The overall height of the nucleotide represents sequence conservation. (c) Aag2 cells transfected with different combinations of dsRNA (all Piwi/Ago 3: Piwi1/2/3, 4, 5, 6, 7 and Ago 3; all Piwi 4: Piwi1/2/3, 5, 6, 7 and Ago 3; Piwi 4 or dsRNA control: eGFP specific) were infected with SFV4(3H)-*RLuc* 24 h p.t. at an m.o.i. of 0.1. The means of three independent experiments performed in triplicate are shown with standard errors (* represents $p<0.05$, *t*-test).

et al., 2012; Morazzani *et al.*, 2012; Schnettler *et al.*, 2013; Scott *et al.*, 2010; Vodovar *et al.*, 2012). However, this role has not been experimentally proven. The piRNA pathway is known to target transposons and thereby ensures genome stability, especially in germline cells. As some arboviruses have been reported to be vertically transmitted (Anderson *et al.*, 2012; Mulyatno *et al.*, 2012), an antiviral response by the piRNA pathway in germline cells may constitute an antiviral mechanism to inhibit vertical transmission or limit virus replication in developing embryos. On the other hand, a putative piRNA pathway in somatic tissues could add another layer to small RNA-based antiviral responses controlling arboviral infection. The finding that SFV-produced piRNA-like small RNA molecules in *Ae. aegypti*- and *Ae. albopictus*-derived cell lines is in accordance with recently published work showing similar results for CHIKV, SINV, LACV and SBV (Morazzani *et al.*, 2012; Schnettler *et al.*, 2013; Vodovar *et al.*, 2012). The observation that the knockdown of some PIWI proteins in Aag2 cells has a positive effect on SFV infection supports the hypothesis that the piRNA pathway and possibly the viral-specific piRNA-like small RNAs have an antiviral function in these cells. A similar result has been previously reported in anopheline mosquitoes. Indeed, *A. gambiae* showed an increase in O'nyong-nyong virus (*Togaviridae; Alphavirus*) following Ago 3 knockdown (Keene *et al.*, 2004). We extend this finding to aedine mosquitoes and highlight the additional contribution of Piwi 4. The fact that virus-specific piRNA-like small RNA molecules are not specific to *Ae. aegypti* but can also be found in infected *Ae. albopictus*, coupled with the expression of all piRNA pathway proteins (PIWIs and Ago 3) in somatic tissues (head

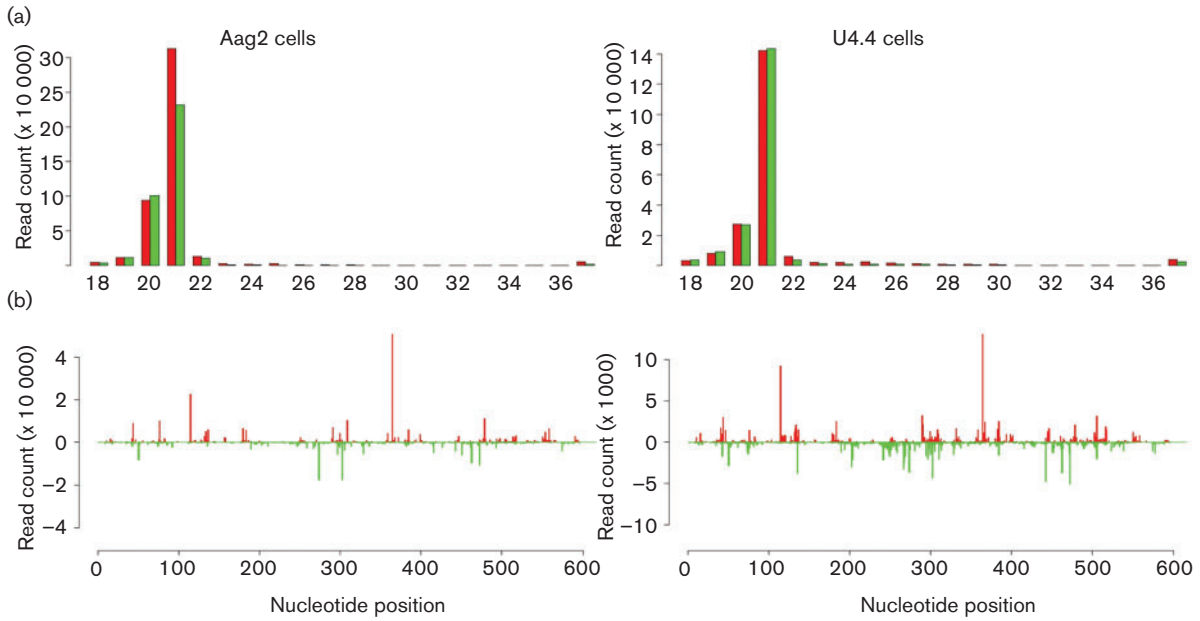


Fig. 5. Transfected dsRNA feeds into the siRNA but not the piRNA pathway. (a) Size distribution of small RNA molecules mapping to the input eGFP sequence following transfection of eGFP-derived dsRNA (720 nt) into *Ae. aegypti* (Aag2) or *Ae. albopictus* (U4.4)-derived cells. RNA was isolated at 24 h p.t.. Red and green colour maps show the coding strand and non-coding strand of the input dsRNA, respectively. (b) Frequency distribution of 21 nt small RNA molecules to the eGFP coding strand (positive) or non-coding (negative) strand in Aag2 and U4.4. The *y*-axis shows the frequency of the 21 nt small RNAs mapping to the corresponding nt position of the 720 nt eGFP-specific dsRNA of the *x*-axis.

and thorax) (Morazzani *et al.*, 2012), indicates that the 'antiviral' piRNA pathway is probably not specific to *Ae. aegypti* but could possibly be present in *Ae. albopictus* as well. It is not known if the same is true for drosophila. Viral-specific piRNAs have been described in drosophila OSS (Wu *et al.*, 2010), but it is not known if they have any antiviral activity in these cells. In addition, no viral-specific piRNAs have been reported in somatic tissue or derived cells of drosophila until now, which is in contrast to aedine mosquitoes and their derived cells (Hess *et al.*, 2011; Léger *et al.*, 2012; Morazzani *et al.*, 2012; Schnettler *et al.*, 2013; Vodovar *et al.*, 2012). This could be due to the differences in PIWI pathway protein expression between drosophila and *Ae. aegypti* (Campbell *et al.*, 2008a). However, knockdown of Piwi in drosophila results in increased WNV production similar to that observed in Ago 2 knockdowns (Chotkowski *et al.*, 2008), which would also support an antiviral activity of the piRNA-related pathway in drosophila. More research is needed to determine the possible antiviral activity of piRNAs in drosophila and whether this is restricted to ovary cells or is found in all somatic tissue, and to determine the precise differences between these pathways in aedine mosquitoes and drosophila.

We do not know how the antiviral piRNA pathway is induced in aedine mosquitoes, although previous observations have suggested a dsRNA molecule as the inducer (Morazzani *et al.*, 2012). This would suggest crosstalk between the siRNA and piRNA pathways. A similar result has been reported for *A. gambiae*-derived cells, which show a decrease in dsRNA-induced reporter gene expression following Ago 3 knockdown (Hoa *et al.*, 2003), indicating such crosstalk even in non-aedine mosquitoes. However, the lack of piRNA-like molecules produced in the case of dsRNA transfection alone (Fig. 5) suggests the need for ssRNA (active viral replication) to induce piRNA production. The inhibitory effect observed by the expression of the dsRNA-binding RNAi suppressor B2 by CHIKV on the production of viral-specific piRNAs suggests that this is a secondary effect as dsRNAs are replication intermediates required for ssRNA production (Morazzani *et al.*, 2012). The observation that most piRNAs map to the coding strand region of the 5′ end of the SFV subgenomic mRNA, SINV or CHIKV (Morazzani *et al.*, 2012; Vodovar *et al.*, 2012) suggests that perhaps particular transcripts or genome regions are preferentially targeted. In this case, viral-specific dsRNA, either due to the sequence or structural features such as dsRNA, could be the inducer of the piRNA pathway. Characterization of the viral-specific piRNA-like molecules suggests a ping-pong production mechanism; however, knockdown of Ago 3, which is known to be important for the ping-pong mechanism in drosophila, did not result in an increase of SFV replication. It could be possible that the observed viral-specific piRNA-like molecules are produced in an Ago 3 independent manner in *Ae. aegypti* in contrast to

drosophila, or that the obtained knockdown of Ago 3 was not sufficient.

To date, it is not definitively known if the viral-specific piRNA-like molecules in mosquitoes and derived cell lines are really produced through the piRNA pathway using PIWI and Ago 3 proteins, although their ping-pong signature highly suggests this production pathway. In addition, piRNA production models were shown in drosophila using at least two PIWI family proteins for the production of primary and secondary piRNA molecules, but knockdown experiments only showed a strong effect on SFV production for Piwi 4. The low frequency of SFV-specific piRNA-like molecules found in cells with knockdown of all Piwi and Ago 3 proteins strongly supports their involvement in the production of these molecules. However, the lack of decrease in SFV-specific piRNA-like molecules in Piwi 4 knockdowns and the increase in SFV replication and production suggest an effector role of this PIWI-clade Ago protein by using the SFV-specific piRNA-like molecules to target the virus. The observed increase in SFV replication in both Piwi 4 and all Piwi/Ago3 knockdown cells compared to control dsRNA could also explain the increase of 21 nt viRNAs in these cells. In addition, the increase in SFV-specific piRNA-like molecules in combination with a higher SFV replication again suggests that Piwi 4 is not needed for SFV-specific piRNA-like production, but rather it is used to target and thereby silence the virus.

Together, these results show that arbovirus replication is able to trigger the piRNA pathway and that silencing of piRNA-related proteins reduces viral-specific piRNA-like molecules and enhances viral replication and production, suggesting an antiviral response by the piRNA pathway. Both the piRNA and exogenous siRNA pathways may act in combination to control viral infections in mosquito cells. Future research is needed to determine the viral inducer molecule of the piRNA pathway and map the involvement of each Piwi/Ago 3 protein in detail. We cannot exclude that some Piwi-clade proteins that are important in viral piRNA-like production have been missed due to either inefficient knockdown or the need of combinational knockdowns, but our results already suggest Piwi 4 as an effector protein. Besides, the proposed linkage between the siRNA and piRNA pathways has yet to be investigated, and it is not yet known if the piRNA and siRNA pathways are restricting different parts of the viral infection (acute versus persistent infection) in mosquitoes. Experiments in the exogenous RNAi pathway knockout cell lines, such as C6/36 (Brackney *et al.*, 2010; Scott *et al.*, 2010), suggest that the piRNA pathway may still be able to control viral infection to some extent on its own; however, further studies are required to fully assess interaction between the pathways.

METHODS

Cells, plasmids and virus. *Ae. albopictus*-derived U4.4 and *Ae. aegypti*-derived Aag2 cells were maintained in L-15 medium supplemented with 10 % FCS and 10 % tryptose phosphate broth at 28 °C. Amplification and titration of SFV (strain SFV4) and the SFV4(3H)-*RLuc* reporter virus and infection of U4.4 and Aag2 cells were performed in a similar way as previously described; infections were performed at growth temperature (28 °C) (Siu *et al.*, 2011). Briefly, viruses were grown in BHK-21 cells in Glasgow minimum essential medium (GMEM) with 5 % FCS and 10 % tryptose phosphate broth at 37 °C with 5 % CO_2. Virus purified from the supernatant or virus present in supernatant was titrated by plaque assay on BHK-21 cells using an Avicell (0.6 %)/MEM overlay with 2 % FCS. Infection of mosquito cells was performed in L-15 medium with 10 % FCS and 10 % tryptose phosphate broth for 1 h at 28 °C, followed by a washing step with PBS and overlay with media.

Reverse transcription and PCR. RT-PCR was performed with total RNA (500 ng) isolated using TRIzol (Invitrogen), Superscript III and oligo-dT primer, according to the manufacturer's protocol. Piwi/Ago 3 transcripts were detected and amplified by PCR (2 µl of the cDNA reaction) using primers containing T7 RNA polymerase promoter sequences (Table S1). For the detection of the transcripts, 40 rounds of PCR using KOD polymerase were performed in contrast to 35 rounds for semi-quantitative PCR using GoTaq polymerase. The eGFP-derived PCR product was produced by using eGFP-C1 (Clontech) as a template. PCR products were gel-purified and used for dsRNA production or first cloned into the pJet blunt 1.2 vector (Fermentas) and sequenced.

***In vitro* dsRNA transcription.** dsRNA molecules for Piwi/Ago 3 and eGFP were produced with a T7 RNA polymerase *in vitro* transcription kit (Megascript RNAi kit, Ambion) using a PCR product as a template, followed by column purification. Internally fluorescently labelled eGFP-specific dsRNA was produced in the same way but using fluorescein-labelled rNTP mix (Roche) following the manufacturer's protocol, and purified by ethanol precipitation. Primer sequences are indicated in Table S1.

Cell viability assay. Viability of cells transfected with dsRNA molecules was determined using a CellTiter-Glo luminescent cell viability assay (Promega) following the manufacturer's recommendations.

Luciferase assay. Luciferase activities were determined using a Dual Luciferase assay kit (Promega) on a GloMax luminometer following cell lysis in Passive Lysis Buffer.

Transfection. Aag2 cells (1.7×10^5 per well) were seeded in 24-well plates, 24 h before transfection. Piwi/Ago 3 transcripts were silenced by the transfection of 100 ng dsRNA per well (Piwi specific or 400 nt eGFP) at 24 h post-seeding with Lipofectamine 2000 (Invitrogen), following the manufacturer's protocol. At 24 h p.t., cells were either harvested to isolate RNA for RT-PCR or infected with SFV4(3H)-*RLuc* at the indicated m.o.i. Supernatant from infected cells (m.o.i. of 0.1) was used to determine virus titre by plaque assays on BHK-21 cells. In addition, luciferase expression was measured 48 h p.i. as described above.

Small RNA isolation and sequencing. Small RNA sequencing was carried out by ARK-Genomics (The Roslin Institute, University of Edinburgh) and The GenePool (University of Edinburgh) using the Illumina Solexa platform. Approximately 5×10^5 U4.4 cells and 6×10^5 Aag2 cells per well were transfected in a 6-well plate with 1 µg eGFP-derived dsRNA (720 nt) or left untreated.

For the infection experiments, Aag2 cells were transfected with 1 µg Piwi 4 or eGFP dsRNA or 200 ng each Piwi1-3, 4, 5, 6, 7 and Ago 3 dsRNA using Lipofectamine 2000. At 24 h p.t., cells were infected with SFV4 at an m.o.i. of 10. At 24 h p.t. or p.i., RNA was isolated using 1 ml TRIzol (Invitrogen) per well, followed by purification,

sequencing and analysis as previously described (Schnettler *et al.*, 2013).

Immunostaining. Aag2 cells were fixed in formaldehyde and permeabilized by 0.3 % Triton/PBS for 30 min, followed by a wash with PBS. Cells were pre-incubated with CAS-Block for 1 h at room temperature, followed by incubation with CAS-Block diluted SFV nsP3-specific antibody (1 : 500) (Siu *et al.*, 2011) for 90 min at room temperature. After three washing steps with PBS, an anti-rabbit antibody conjugated with Alexa Fluor 543 diluted in CAS-Block (1 : 1000) was incubated for 60 min at room temperature. Following further washing steps with PBS, cells were dried and mounted with DAPI-containing hard set Vectashield mounting medium (Vector Laboratories), and fluorescence was detected on a Zeiss LSM Meta microscope.

ACKNOWLEDGEMENTS

This work was supported by a Rubicon fellowship (Netherlands Organisation for Scientific Research NWO, grant 825.10.021) (E. S.), a Southern African Research fellowship (S. H.), the UK Medical Research Council (A. K.), a BBSRC Roslin Institute Strategic Programme Grant (J. K. F., A. K.) and The Pirbright Institute (J. K. F., R. F.). We thank Andres Merits for providing viral constructs.

REFERENCES

Anderson, J. F., Main, A. J., Cheng, G., Ferrandino, F. J. & Fikrig, E. (2012). Horizontal and vertical transmission of West Nile virus genotype NY99 by *Culex salinarius* and genotypes NY99 and WN02 by *Culex tarsalis*. *Am J Trop Med Hyg* **86**, 134–139.

Blair, C. D. (2011). Mosquito RNAi is the major innate immune pathway controlling arbovirus infection and transmission. *Future Microbiol* **6**, 265–277.

Brackney, D. E., Scott, J. C., Sagawa, F., Woodward, J. E., Miller, N. A., Schilkey, F. D., Mudge, J., Wilusz, J., Olson, K. E. & other authors (2010). C6/36 *Aedes albopictus* cells have a dysfunctional antiviral RNA interference response. *PLoS Negl Trop Dis* **4**, e856.

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R. & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103.

Campbell, C. L., Black, W. C., IV, Hess, A. M. & Foy, B. D. (2008a). Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. *BMC Genomics* **9**, 425.

Campbell, C. L., Keene, K. M., Brackney, D. E., Olson, K. E., Blair, C. D., Wilusz, J. & Foy, B. D. (2008b). *Aedes aegypti* uses RNA interference in defense against Sindbis virus infection. *BMC Microbiol* **8**, 47.

Chotkowski, H. L., Ciota, A. T., Jia, Y., Puig-Basagoiti, F., Kramer, L. D., Shi, P. Y. & Glaser, R. L. (2008). West Nile virus infection of *Drosophila melanogaster* induces a protective RNAi response. *Virology* **377**, 197–206.

Donald, C. L., Kohl, A. & Schnettler, E. (2012). New insights into control of arbovirus replication and spread by insect RNA interference pathways. *Insects* **3**, 511–531.

Hess, A. M., Prasad, A. N., Ptitsyn, A., Ebel, G. D., Olson, K. E., Barbacioru, C., Monighetti, C. & Campbell, C. L. (2011). Small RNA profiling of Dengue virus-mosquito interactions implicates the PIWI RNA pathway in anti-viral defense. *BMC Microbiol* **11**, 45.

Hoa, N. T., Keene, K. M., Olson, K. E. & Zheng, L. (2003). Characterization of RNA interference in an *Anopheles gambiae* cell line. *Insect Biochem Mol Biol* **33**, 949–957.

Keene, K. M., Foy, B. D., Sanchez-Vargas, I., Beaty, B. J., Blair, C. D. & Olson, K. E. (2004). RNA interference acts as a natural antiviral response to O'nyong-nyong virus (Alphavirus; Togaviridae) infection of *Anopheles gambiae*. *Proc Natl Acad Sci U S A* **101**, 17240–17245.

Léger, P., Lara, E., Jagla, B., Sismeiro, O., Mansuroglu, Z., Coppée, J. Y., Bonnefoy, E. & Bouloy, M. (2013). Dicer-2 and Piwi mediated RNA interference in Rift Valley Fever virus infected mosquito cells. *J Virol* **87**, 1631–1648.

Morazzani, E. M., Wiley, M. R., Murreddu, M. G., Adelman, Z. N. & Myles, K. M. (2012). Production of virus-derived ping-pong-dependent piRNA-like small RNAs in the mosquito soma. *PLoS Pathog* **8**, e1002470.

Mulyatno, K. C., Yamanaka, A., Yotopranoto, S. & Konishi, E. (2012). Vertical transmission of dengue virus in *Aedes aegypti* collected in Surabaya, Indonesia, during 2008-2011. *Jpn J Infect Dis* **65**, 274–276.

Saito, K. & Siomi, M. C. (2010). Small RNA-mediated quiescence of transposable elements in animals. *Dev Cell* **19**, 687–697.

Sánchez-Vargas, I., Scott, J. C., Poole-Smith, B. K., Franz, A. W., Barbosa-Solomieu, V., Wilusz, J., Olson, K. E. & Blair, C. D. (2009). Dengue virus type 2 infections of *Aedes aegypti* are modulated by the mosquito's RNA interference pathway. *PLoS Pathog* **5**, e1000299.

Schnettler, E., Ratinier, M., Watson, M., Shaw, A. E., McFarlane, M., Varela, M., Elliott, R. M., Palmarini, M. & Kohl, A. (2013). RNA interference targets arbovirus replication in *Culicoides* cells. *J Virol* **87**, 2441–2454.

Scott, J. C., Brackney, D. E., Campbell, C. L., Bondu-Hawkins, V., Hjelle, B., Ebel, G. D., Olson, K. E. & Blair, C. D. (2010). Comparison of dengue virus type 2-specific small RNAs from RNA interference-competent and -incompetent mosquito cells. *PLoS Negl Trop Dis* **4**, e848.

Senti, K. A. & Brennecke, J. (2010). The piRNA pathway: a fly's perspective on the guardian of the genome. *Trends Genet* **26**, 499–509.

Siomi, M. C., Miyoshi, T. & Siomi, H. (2010). piRNA-mediated silencing in *Drosophila* germlines. *Semin Cell Dev Biol* **21**, 754–759.

Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* **12**, 246–258.

Siu, R. W., Fragkoudis, R., Simmonds, P., Donald, C. L., Chase-Topping, M. E., Barry, G., Attarzadeh-Yazdi, G., Rodriguez-Andres, J., Nash, A. A. & other authors (2011). Antiviral RNA interference responses induced by Semliki Forest virus infection of mosquito cells: characterization, origin, and frequency-dependent functions of virus-derived small interfering RNAs. *J Virol* **85**, 2907–2917.

van Rij, R. P. & Berezikov, E. (2009). Small RNAs and the control of transposons and viruses in *Drosophila*. *Trends Microbiol* **17**, 163–171.

van Rij, R. P., Saleh, M. C., Berry, B., Foo, C., Houk, A., Antoniewski, C. & Andino, R. (2006). The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in *Drosophila melanogaster*. *Genes Dev* **20**, 2985–2995.

Vodovar, N., Bronkhorst, A. W., van Cleef, K. W., Miesen, P., Blanc, H., van Rij, R. P. & Saleh, M. C. (2012). Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. *PLoS ONE* **7**, e30861.

Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E. C., Li, W. X. & Ding, S. W. (2010). Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc Natl Acad Sci U S A* **107**, 1606–1611.

Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways

Nicholas R. Thomson, Debra J. Clayton, Daniel Windhorst, et al.

Genome Res. 2008 18: 1624-1637 originally published online June 26, 2008
Access the most recent version at doi:10.1101/gr.077404.108

Letter

Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways

Nicholas R. Thomson,[1,9] Debra J. Clayton,[2] Daniel Windhorst,[3] Georgios Vernikos,[1] Susanne Davidson,[2] Carol Churcher,[1] Michael A. Quail,[1] Mark Stevens,[2] Michael A. Jones,[4] Michael Watson,[2] Andy Barron,[1] Abigail Layton,[2] Derek Pickard,[1] Robert A. Kingsley,[1] Alex Bignell,[1] Louise Clark,[1] Barbara Harris,[1] Doug Ormond,[1] Zahra Abdellah,[1] Karen Brooks,[1] Inna Cherevach,[1] Tracey Chillingworth,[1] John Woodward,[1] Halina Norberczak,[1] Angela Lord,[1] Claire Arrowsmith,[1] Kay Jagels,[1] Sharon Moule,[1] Karen Mungall,[1] Mandy Sanders,[1] Sally Whitehead,[1] Jose A. Chabalgoity,[5] Duncan Maskell,[6] Tom Humphrey,[7] Mark Roberts,[8] Paul A. Barrow,[4] Gordon Dougan,[1] and Julian Parkhill[1]

[1]The Pathogen Sequencing Unit, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; [2]Division of Microbiology, Institute for Animal Health, Compton, Berkshire RG20 7NN, United Kingdom; [3]Lohmann Animal Health GmbH & Co. KG, 27472 Cuxhaven, Germany; [4]School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington, Leicestershire LE12 5RD, United Kingdom; [5]Department of Biotechnology, School of Medicine, Universidad de la Republica, Montevideo CP 11600, Uruguay; [6]Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, United Kingdom; [7]School of Clinical Veterinary Science, University of Bristol, Langford, Bristol BS40 5DU, United Kingdom; [8]Institute of Comparative Medicine, Faculty of Veterinary Medicine, University of Glasgow, Glasgow G61 1QH, United Kingdom

We have determined the complete genome sequences of a host-promiscuous *Salmonella enterica* serovar Enteritidis PT4 isolate P125109 and a chicken-restricted *Salmonella enterica* serovar Gallinarum isolate 287/91. Genome comparisons between these and other *Salmonella* isolates indicate that *S.* Gallinarum 287/91 is a recently evolved descendent of *S.* Enteritidis. Significantly, the genome of *S.* Gallinarum has undergone extensive degradation through deletion and pseudogene formation. Comparison of the pseudogenes in *S.* Gallinarum with those identified previously in other host-adapted bacteria reveals the loss of many common functional traits and provides insights into possible mechanisms of host and tissue adaptation. We propose that experimental analysis in chickens and mice of *S.* Enteritidis–harboring mutations in functional homologs of the pseudogenes present in *S.* Gallinarum could provide an experimentally tractable route toward unraveling the genetic basis of host adaptation in *S. enterica*.

[Supplemental material is available online at www.genome.org. The genome sequence data from this study have been submitted to EMBL under accession nos. AM933172 and AM933173.]

Zoonotic pathogens, particularly those associated with veterinary animals in the human food chain, are some of the most important causes of infectious diseases in humans. Pathogens associated with zoonotic infections exhibit a promiscuous phenotype in that they maintain the ability to colonize and potentially cause infections in more than one host species. In contrast, some pathogenic agents are significantly host restricted, or adapted, and are normally only able to cause disease in one host. *Salmonella enterica* is a single bacterial species that includes examples of both promiscuous and host-adapted pathotypes. Isolates from serovars such as *S. enterica* serovar Typhimurium and *S.* Enteritidis predominantly retain the ability to infect more than one mammalian host, including humans, whereas serovars such as *S. enterica* serovars Typhi and *S.* Gallinarum are restricted to humans and chickens, respectively. The ability to transmit between and within particular host populations is centrally important in dictating the epidemiology of infections and the emergence of new diseases.

Before the mid-1980s, *S.* Enteritidis was regarded as an *S. enterica* serovar of minor public health significance, but subsequently this serovar became dominant in terms of human food poisoning in many parts of the world (Rodrigue et al. 1990). National and international legislation regarding the reporting of disease incidence, improved hygiene and biosecurity (Barrow

[9]Corresponding author.
E-mail nrt@sanger.ac.uk; fax 44-(0)-1223-494919.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.077404.108.

18:1624–1637 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

2000), and vaccination have contributed to controlling *S.* Enteritidis levels in poultry and consequently in man in Europe, but levels of infection remain significant. Most recent isolates of *S.* Enteritidis are regarded as promiscuous in the sense that they can cause infections in mice, retain the ability to colonize the tissues of chickens, and cause gastroenteritis in man.

S. Gallinarum, the causative agent of fowl typhoid, is a predominantly avian-restricted serovar (Shivaprasad 2000). Interestingly, in common with the human-restricted serovar *S.* Typhi, the chicken-adapted *S.* Gallinarum causes an invasive typhoid-like disease. Thus, here host adaptation appears to have coevolved with loss of the intestinal lifestyle and the acquisition of the ability to cause systemic infection. *S.* Gallinarum still causes a disease of worldwide economic significance, and although it has been largely controlled in countries with strong health control policies, largely through serology-based test and slaughter schemes, it remains a problem elsewhere. Multi locus enzyme electrophoresis analyses of isolates of *S.* Enteritidis and *S.* Gallinarum indicate that, together with isolates of *S.* Dublin and *S.* Pullorum, they form a related strain cluster that share the same lipopolysaccharide-based O structure (O-1, 9, 12 characteristic of serogroup D). The nonmotile *S.* Gallinarum and *S.* Pullorum were previously suggested to have split independently from a motile ancestor related to *S.* Enteritidis (Li et al. 1993; McMeechan et al. 2005). Nonmotility in *S.* Gallinarum has been partially attributed to mutations in the flagellin subunit gene *fliC* gene (Kilger and Grimont 1993), which would normally express the phase 1 g, m antigens characteristic of *S.* Enteritidis. Nonmotility may enhance the ability to invade systemically from the gut by avoiding the TLR-5-induced pro-inflammatory responses of the host (Kaiser et al. 2000; Iqbal et al. 2005).

Here we report the full genome sequences of representative isolates of *S.* Enteritidis and *S.* Gallinarum and provide a detailed comparative genomic analysis of the two serovars. These data have been used to provide insight into the biology, mechanisms of host/tissue adaptation, and evolutionary relationships of these important pathogens.

Results and Discussion

General features of the *S.* Enteritidis PT4 strain P125109 and *S.* Gallinarum strain 287/91 genomes

The complete genome sequences of the promiscuous *S.* Enteritidis PT4 strain P125109 (hereafter *S.* Enteritidis PT4; EMBL accession no. AM933172) and the highly host-adapted chicken pathogen *S.* Gallinarum strain 287/91 (hereafter *S.* Gallinarum 287/91; EMBL accession no. AM933173) were determined and annotated. The main features are summarized in Table 1 and Figure 1, where they are compared with *S.* Typhimurium strain LT2 (hereafter *S.* Typhimurium LT2) (McClelland et al. 2001). The most striking feature of the analysis is the predominant similarity and synteny of core regions of the genomes, including many of the *Salmonella* pathogenicity islands (SPI). Indeed, this comparative analysis highlights an extremely close relationship between the genomes of *S.* Enteritidis and *S.* Gallinarum, suggesting the latter is a direct evolutionary descendent of the former. However, in comparison to *S.* Enteritidis PT4, *S.* Gallinarum 287/91 harbors a significantly higher number of predicted pseudogenes. Although the number of pseudogenes in *S.* Enteritidis PT4 is slightly higher than reported for *S.* Typhimurium LT2, it is in line with levels described in other broad host range enteric pathogens such as *Yersinia enterocolitica* (Thomson et al. 2006). In contrast, the number of pseudogenes in *S.* Gallinarum 287/91 is closer to that of the human-restricted *S.* Typhi CT18 (204 pseudogenes) (Parkhill et al. 2001) and *S. enterica* serovar Paratyphi A (173 pseudogenes) (McClelland et al. 2004).

Whole-genome comparisons of *S.* Enteritidis PT4 and *S.* Typhimurium LT2

Initially, the genome of *S.* Enteritidis PT4 was compared with that of *S.* Typhimurium LT2, a well-characterized and fully sequenced *S. enterica* isolate. *S.* Enteritidis PT4 and *S.* Typhimurium LT2 are both representatives of serovars able to cause enteritis in a broad range of hosts and produce murine typhoid, but they also show significant phenotypic differences, including serovar type. An alignment of the genome of *S.* Enteritidis PT4 with that of *S.* Typhimurium LT2 revealed colinearity except for an inversion about the terminus in *S.* Typhimurium LT2 (Fig. 1) (McClelland et al. 2001), with >90% of coding sequences (CDS) forming an extensive core gene-set (Figs. 2, 3). The average nucleotide identity between the shared orthologs is 98.98% compared with 99.7% between those of LT2 and a second fully sequenced *S.* Typhimurium strain SL1344 (data not shown). The genes that are only present either in *S.* Enteritidis PT4 or *S.* Typhimurium LT2 form 6.4% and 9.6% of their respective genomes (Fig. 3). The majority of *S.* Enteritidis PT4 unique CDS are in clusters from >3 kb up to >40 kb, but there are very few indels of <3 kb (Fig. 2; Table 2). We refer to these nonshared gene clusters as regions of difference (ROD). CDS present in *S.* Enteritidis PT4 but absent from *S.* Typhimurium LT2 are dominated by prophage-related functions, although other functional classes are represented (Fig. 3).

Table 1. General properties of *S. enterica* serovar genomes

	S. enterica serovars			
Serovar	**Enteritidis**	**Typhimurium**	**Gallinarum**	**Typhi**
Strain	P125109 (PT4)	LT2	287/91	CT18
Size	4,685,848	4,857,432	4,658,697	4,809,037
Percent G + C content (%)	52.17	52.22	52.20	52.09
No. of CDS	4318	4451	4274	4599
Coding density	85.5%	86.8%	79.9%	87.6%
Average gene size	953	947	939	958
rRNA operons	7	7	7	7
tRNA	84	85	75	78
Pseudogenes[a]	113	25	309	204

[a]Taken from original publications (see text).

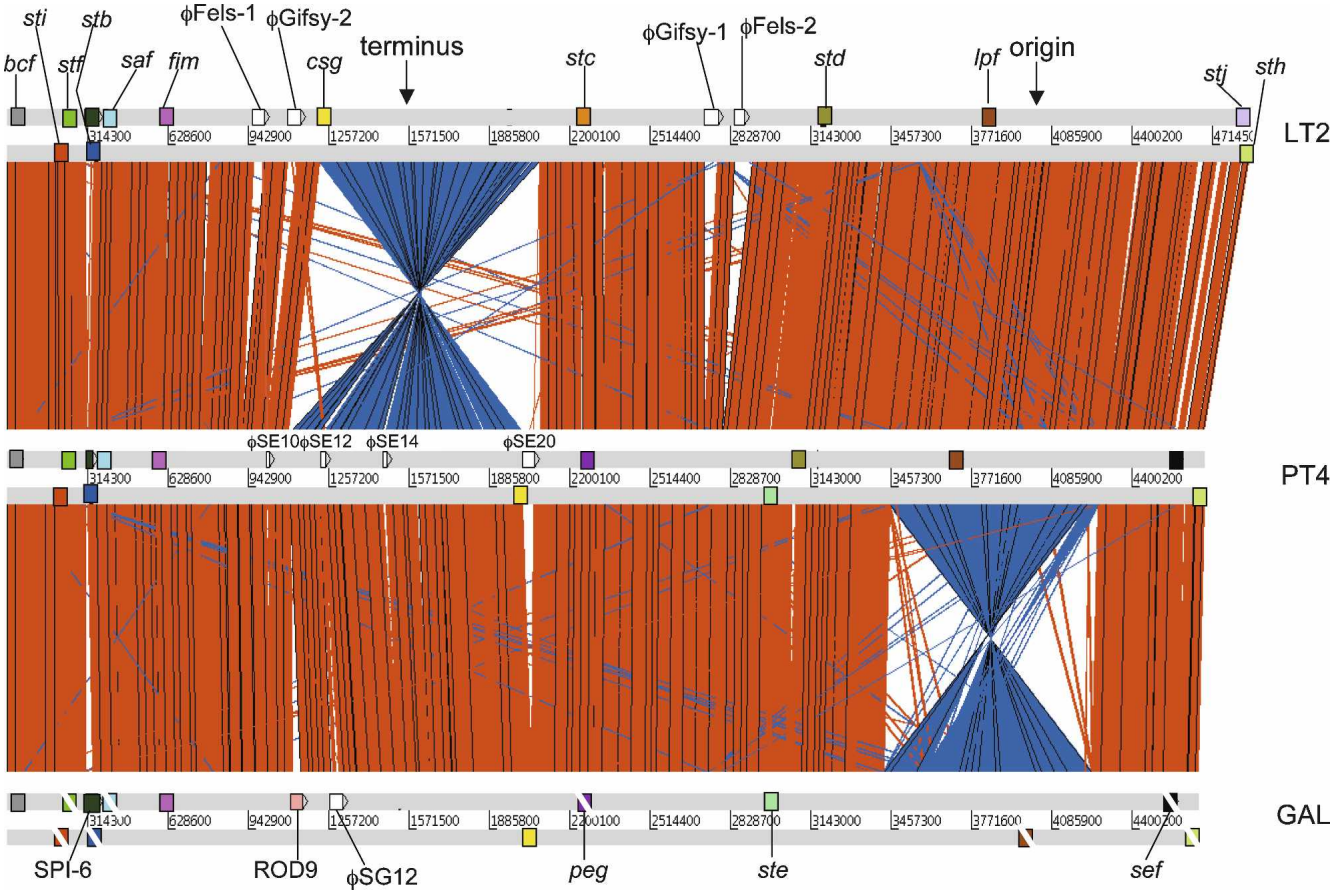


Figure 1. Global comparison between *S.* Typhimurium, *S.* Enteritidis, and *S.* Gallinarum. ACT comparison (http://www.sanger.ac.uk/Software/ACT) of amino acid matches between the complete six-frame translations (computed using TBLASTX) of the whole-genome sequences of *S.* Typhimurium LT2 (LT2), *S.* Enteritidis PT4 (PT4), and *S.* Gallinarum 287/91 (GAL). Forward and reverse strands of DNA are shown for each genome (light gray horizontal bars). The red bars between the DNA lines represent individual TBLASTX matches, with inverted matches colored blue. The position of all the fimbrial operons in these three genomes are marked as colored boxes positioned on the forward and reverse strands of DNA. Analogous fimbrial operons are colored the same. The boxes of fimbrial operons that include pseudogenes are crossed with a white line. Other genomic features are only shown if they constitute breaks in synteny between genomes. The position of the origin and terminus are marked (solid black arrows).

Gene sets common to both *S.* Enteritidis PT4 and *S.* Typhimurium LT2

Within the core genes, there are many of the functions associated with virulence and host interactions and include SPIs and fimbrial operons. With the exception of *SPI-6*, *SPI-9*, and *SPI-10*, the other SPIs in *S.* Enteritidis PT4 are closely related to their equivalents in *S.* Typhimurium LT2 (Fig. 2) (McClelland et al. 2001). Of the three SPIs that vary, *S.* Enteritidis PT4 *SPI-10* only encodes the *sef* fimbrial operon, consistent with this region being mosaic in isolates of different serovars (Edwards et al. 2000; Collighan and Woodward 2001; Bishop et al. 2005). The *SPI-9* CDS *SE2609*, encoding a large repetitive exported protein, appears intact in *S.* Enteritidis PT4 unlike the ortholog, *STM2689*, in *S.* Typhimurium LT2. The *SPI-6* region of *S.* Enteritidis PT4 is 22 kb in size compared with 47 kb in *S.* Typhimurium LT2. *SPI-6* varies markedly in size in all the other sequenced *Salmonella*, including *S.* Typhi (McClelland et al. 2001, 2004; Parkhill et al. 2001; Chiu et al. 2005). Of the other known SPIs, *SPI-8*, *SPI-7*, and *SPI-15* are absent from *S.* Enteritidis PT4 (Parkhill et al. 2001; Vernikos and Parkhill 2006). Conversely, *SPI-17* is present in *S.* Enteritidis PT4 but absent from *S.* Typhimurium LT2. The *S.* Enteritidis PT4 *SPI-17* is a degenerate prophage encoding CDS known to be involved in O-antigen conversion in other systems (Vernikos and Parkhill 2006).

S. Enteritidis PT4 harbors 13 fimbrial clusters, 10 of which are highly conserved in *S.* Typhimurium LT2 with orthologous genes sharing >97% nucleotide identity and inserted at the same sites in both genomes (Fig. 1; Table 2). The only exceptions to this are *safA*, *safB*, and *stdA*, where the *S.* Enteritidis PT4 and *S.* Typhimurium LT2 orthologs show 81%, 87%, and 89% nucleotide identity, respectively. The *S.* Enteritidis PT4 fimbrial clusters not found in *S.* Typhimurium LT2 include a novel cluster we have termed *peg*, which is inserted at the same location as the *S.* Typhimurium LT2 *stc* operon and is so far restricted to *S.* Enteritidis, *S.* Gallinarum 287/91, and *S.* Paratyphi A. The *peg* fimbrial proteins show 58%–64% identity with their predicted functional equivalents in the *S.* Typhimurium LT2 *stc* cluster (Table 2). Of the remaining fimbrial clusters, *ste* is absent from *S.* Typhimurium LT2, but there is a deletion remnant of the *ste* major pilin subunit remaining at the analogous site (*S.* Typhimurium LT2; positions 3,102,016–3,102,150 bps). Fimbrial operon *stj* is present in *S.* Typhimurium LT2 and replaces a gene of unknown function still present in *S.* Enteritidis PT4 (*SEN4331A*). Thus, in common

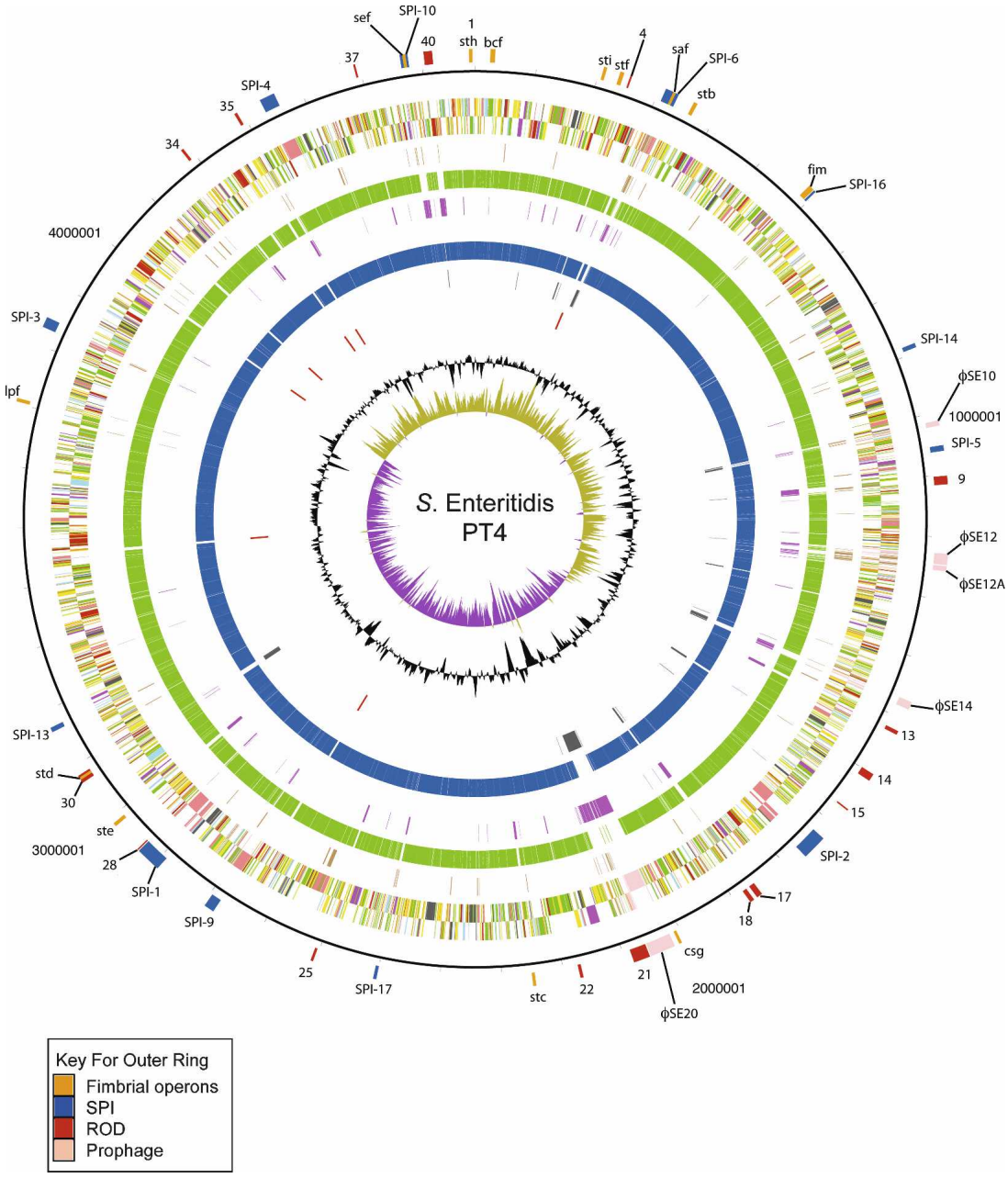


Figure 2. Circular representation of the *S.* Enteritidis PT4 chromosome. From the outside in, the *outer* circle 1 marks the position of regions of difference (mentioned in the text) and is detailed in Table 2. Circle 2 shows the size in base pairs. Circles 3 and 4 show the position of CDS transcribed in a clockwise and anti-clockwise direction, respectively (for color codes see below); circle 5 shows the position of *S.* Enteritidis PT4 pseudogenes. Circles 6 and 8 show the position of *S.* Enteritidis PT4 genes that have orthologs (by reciprocal FASTA analysis) in *S.* Typhimurium strain LT2 (all CDS colored green) and *S.* Gallinarum strain 287/91 (all CDS colored blue), respectively. Circles 7 and 9 show the position of *S.* Enteritidis PT4 genes that lack orthologs in (by reciprocal FASTA analysis) in *S.* Typhimurium strain LT2 (all CDS colored pink) and *S.* Gallinarum strain 287/91 (all CDS colored gray), respectively. Circle 10 shows the position of *S.* Enteritidis PT4 rRNA operons (red). Circle 11 shows a plot of G + C content (in a 10-kb window). Circle 12 shows a plot of GC skew ([G − C]/[G + C]; in a 10-kb window). Genes in circles 3 and 4 are color-coded according to the function of their gene products: dark green, membrane or surface structures; yellow, central or intermediary metabolism; cyan, degradation of macromolecules; red, information transfer/cell division; cerise, degradation of small molecules; pale blue, regulators; salmon pink, pathogenicity or adaptation; black, energy metabolism; orange, conserved hypothetical; pale green, unknown; and brown, pseudogenes.

with other promiscuous salmonellae, *S.* Enteritidis PT4 harbors multiple functional fimbrial operons (Townsend et al. 2001).

In addition to the gene remnants found for the *ste* fimbrial operon *S.* Enteritidis PT4, RODs *ROD17*, *ROD25*, *ROD34*, *ROD35*, and *ROD37* (Table 2) that fall outside of the core gene-set also have discernable remnants in *S.* Typhimurium LT2 or are conserved in other *S. enterica* and so are likely to have been present in a precursor of *S.* Enteritidis PT4, shared with *S.* Typhimurium LT2, and subsequently deleted from *S.* Typhimurium LT2.

Gene sets only present in *S.* Enteritidis PT4 but not *S.* Typhimurium LT2

By analyzing the genetic context of the *S.* Enteritidis PT4 specific RODs, we distinguish between those likely to have been acquired independently from those that may have been deleted from *S.* Typhimurium LT2 (discussed above). Examples of likely acquisitions include *SPI-17*, *ROD4*, *ROD9*, *ROD13*, *ROD21*, *ROD22*, *ROD28*, *ROD40*, ϕSE14, and ϕSE20 (Table 2). RODs unique to *S.* Enteritidis PT4 include potentially mobile genomic islands, clusters of genes encoding metabolic functions and prophage-like elements, as well as a variable assortment of fimbrial operons already described (summarized in Table 2).

Genomic islands

ROD21 is the only *S.* Enteritidis PT4 genomic island not found in *S.* Typhimurium LT2 and has features characteristic of mobile genetic elements (Table 2). *ROD21* shares significant structural conservation with conserved genomic loci present in a range of bacteria. These islands all display an unusual G + C profile, whereby regions conserved between islands show a higher G + C content compared with the variable island-specific regions (Fig. 4) (Williamson and Free 2005; this study). Surprisingly, most of the *ROD21*-related islands encode paralogs of H-NS (*hnsB*) and/or an H-NS antagonist, *hnsT* (Williamson and Free 2005; Navarre et al. 2006; Doyle et al. 2007). These paralogs may play a role in relieving any potential fitness burden associated with sequestering H-NS to these low G + C DNA elements (Doyle et al. 2007).

Of the other *S.* Enteritidis PT4-specific RODs, *ROD13* encodes five CDS displaying sequence similarities and synteny with genes associated with the uptake and catabolism of the hexonate sugar acid *L*-idonate encoded by the *gntII* locus of *Escherichia coli* (Table 2) (Bausch et al. 1998). Although the substrate for this system is unclear, it is known that colonic mucus contains several sugar acids that represent an important source of nutrients and that *E. coli* mutants unable to utilize them are unable to colonize the mouse large intestine (Sweeney et al. 1996). Moreover, genes involved in the transport of gluconate and related hexonates are up-regulated in *S.* Typhimurium in macrophage, suggesting that they may also be an important source of carbon for intracellular bacteria (Eriksson et al. 2003).

The *S.* Enteritidis PT4 RODs also include loci that are highly variable in the salmonellae and, for *ROD40*, between the wider *Enterobacteriaceae*. *ROD40* locus encodes a Type I restriction/modification system and is analogous to the variable *E. coli* immigration control region (ICR) (Raleigh 1992; Titheradge et al. 1996).

Prophage

Prophage are known to drive diversity in *S. enterica*, and thus, it is not surprising that many *S.* Enteritidis PT4–specific RODs are prophage-like elements, including ϕSE10, ϕSE12, ϕSE12A, ϕSE14, and ϕSE20 (Fig. 2; Table 2) (Thomson et al. 2004; Cooke et al. 2007). All these prophage regions are related and carry the same cargo genes as prophage found previously in other *S. enterica* (see Table 2), including genes encoding type three secretion system (TTSS) effector proteins—*sseK3*, *sspH2*, *gogA*, *sseI*, and *sopE*; the PhoPQ-activated genes *pagK* and *pagM*; as well as *sodCI* encoding a Cu/Zn superoxide dismutase known to be an important colonization factor for *S.* Typhimurium (Stanley et al. 2000; Figueroa-Bossi et al. 2001; Mmolawa et al. 2003; Thomson et al. 2004). Of the six prophage-related regions, only ϕSE20 appears intact and probably represents a recent insertion event, whereas remnants of ϕSE12A are also present at the same location in *S.* Typhimurium LT2 and probably represent the most ancient phage insertion that has been maintained in these two *Salmonella* lineages. However, the number of remnants and intact cargo genes found on the *S.* Enteritidis PT4 prophage highlights the importance of these elements for gene sampling and increasing the overall diversity and even pathogenic potential of salmonellae.

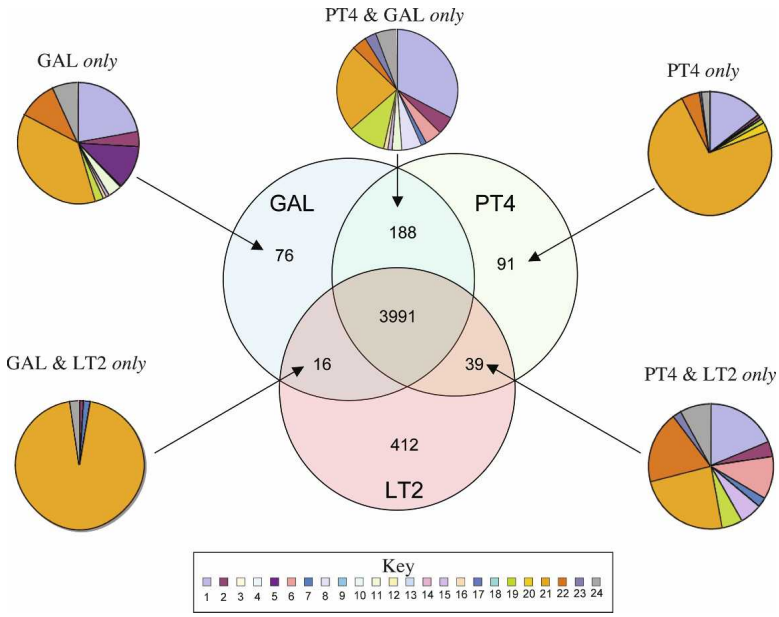


Figure 3. Distribution of orthologous CDS in *S.* Enteritidis PT4, *S.* Typhimurium LT2, and *S.* Gallinarum 287/91. The Venn diagram shows the number of genes unique or shared between two other *S. enterica* serovars (see Methods). The associated pie charts show the breakdown of the functional groups assigned for CDS in relevant sections of the Venn diagram. Color code for the pie charts is as follows: (1) hypothetical proteins, (2) conserved hypothetical proteins, (3) chemotaxis and motility, (4) chromosomal replication, (5) chaperones, (6) protective responses, (7) transport and binding proteins, (8) adaptations to atypical conditions, (9) cell division, (10) macromolecule degradation, (11) synthesis and modification of macromolecules, (12) amino acid biosynthesis, (13) biosynthesis of cofactors, prosthetic groups and carriers, (14) central intermediary metabolism, (15) small-molecule degradation, (16) energy metabolism, (17) fatty acid biosynthesis, (18) nucleosides and nucleotide biosynthesis and metabolism, (19) periplasmic/exported/lipoproteins, (20) ribosomal proteins, (21) laterally acquired (including prophage CDS), (22) pathogenicity and virulence, (23) general regulation, and (24) miscellaneous function. PT4 indicates *S.* Enteritidis PT4; LT2, *S.* Typhimurium LT2; and GAL, *S.* Gallinarum 287/91

Whole-genome sequence of *S.* Gallinarum 287/91 and comparisons with *S.* Enteritidis PT4 and *S.* Typhimurium LT2

One of the striking features of the *S.* Gallinarum 287/91 genome is the high simi

Table 2. The *S.* Enteritidis PT4 variable genome regions identified by genome sequencing

Label[a]	PT4 CDS range	GAL CDS range	Locus name(s)[b]	Size in PT4[c]	General description of locus	Where present
bcf	*SEN0020–SEN0027*	*SG0023–SG0030*	*bcf*	7.6 kb	Fimbrial operon (bovine colonization factor)	GAL PT4 LT2
sti	*SEN0179–SEN0182*	*SG0177–SG0180*	*sti*	4.9 kb	Fimbrial operon (*S.* Typhi I)	GAL PT4 LT2
stf	*SEN0200–SEN0205*	*SG0199–SG0204*	*stf*	6.4 kb	Fimbrial operon (*S.* Typhi F)	GAL PT4 LT2
4	*SEN0216*	*NP*	*ROD4*	3 kb	Viral enhancing factor (metalloprotease)	PT4, NI LT2, or GAL
SPI-6	*SEN0267–SEN0290*	*SG0263–SG0318*	*SPI-6*	17.6 kb (44 kb)	*Salmonella* pathogenicity island (tRNA-*asp*)	GAL PT4 LT2. Variable in all three (Parkhill et al. 2001)
saf	*SEN0281–SEN0284*	*SG0308–SG0312*	*saf*	4.3 kb	Part of SPI-6. Fimbriae: *Salmonella* atypical fimbriae	GAL PT4 LT2
stb	*SEN0319–SEN0323*	*SG0346–SG0350*	*stb*	5.9 kb	Fimbrial operon (*S.* Typhi B)	GAL PT4 LT2
fim	*SEN0524–SEN0533*	*SG0555–SG0564*	*fim*	9 kb	Fimbrial operon (tRNA-*arg*)	GAL PT4 LT2
SPI-16	*SEN0535–SEN0537*	*SG0567–SG0569*	*SPI-16*	3.3 kb	Phage remnant. Cargo: LPS modification genes (tRNA-*arg*)	GAL PT4 LT2 (Vernikos and Parkhill 2006)
SPI-14	*SEN0800–SEN0805*	*SG0835–SG0840*	*SPI-14*	6.8 kb	Electron transfer and regulatory CDS	GAL PT4 LT2 (Shah et al. 2005)
ϕSE10	*SEN0908A–SEN0921*	*NP*	ϕSE10	8.2 kb	Prophage (remnant). Related to ϕGifsy-2. Cargo: *sseI*, *gtgE*, *gtgF*	PT4 LT2, NI: GAL
SPI-5	*SEN0951–SEN0958*	*SG0976–SG0985*	*SPI-5*	6.6 kb	*Salmonella* pathogenicity island (tRNA-*ser*)	GAL PT4 LT2 (Wood et al. 1998)
9	*SEN0995–SEN1013B*	*SG1023–SG1058*	*ROD9*	13.7 kb (42 kb)	RHS element, exported proteins and an IcmF-like CDS	GAL. NI: LT2. Degenerate in PT4
ϕSE12	*SEN1131–SEN1156*	*SG1180–SG1234*	ϕSE12 [ϕSG12]	18 kb (45 kb)	Prophage (remnant). Related to ϕGifsy-2 Cargo: *sodCI*, *ompX**, *sopE*, *gogA**, *hokW*	GAL. NI: LT2. Degenerate in PT4
ϕSE12A	*SEN1158–SEN1171B*	*SG1236–SG1249*	ϕSE12A [ϕSG12A]	8 kb	Prophage (remnant). Related to ϕGifsy-2 Cargo: *pagK*, *pagM**	GAL PT4 LT2
ϕSE14	*SEN1378–SEN1398*	*NP*	ϕSE14	12.6 kb	Prophage (remnant). Related to *S.* Typhi ϕST18	PT4, NI: LT2 GAL
13	*SEN1432–SEN1436*	*SG1499–SG1503*	*ROD13*	6 kb	Possible hexonate uptake and catabolism operon	GAL PT4, NI LT2
14	*SEN1499–SEN1509*	*SG1573*–SG1576**	*ROD14*	11.7 kb (1.8 kb)	Drug efflux system, *pqaA*	GAL PT4 LT2. Variable in all three
15	*SEN1565–SEN1567*	*SG1636**	*ROD15*	2.2 kb (0.6 kb)	Exported protease, *mdtI* and *mdtJ* (Nal^r, Fos^r, Det^r) (Nishino and Yamaguchi 2001)	PT4 LT2. Remnant in GAL
SPI-2	*SEN1623–SEN1666*	*SG1694–SG1738*	*SPI-2*	39.8 kb	*Salmonella* pathogenicity island (tRNA-*val*)	PT4 GAL LT2 (Hensel 2000)
17	*SEN1751–SEN1758*	*SG1824–SG1832*	*ROD17*	9.3 kb	Putative membrane-transport system	PT4 GAL. Remnant in LT2 (1,370,980–1,371,012)
18	*SEN1766–SEN1769*	*SG1841**	*ROD18*	6.9 kb (0.3 kb)	*mipA*—outer-membrane scaffold	PT4 LT2. Remnant in GAL
csg	*SEN1903–SEN1909*	*SG1977–SG1983*	*csg*	4.4 kb	Curli fimbrial operon	PT4 GAL LT2
ϕSE20	*SEN1919A–SEN1966*	*NP*	ϕSE20	40.6 kb	Prophage. Related to ϕST64B. Cargo: *sopE**, *sspH1**, *orgA* (tRNA-*ser*)	PT4, NI: GAL LT2
21	*SEN1970–SEN1999*	*SG1996–SG2025*	*ROD21*	26.5 kb	Genomic island (61 bp DR; tRNA-*asn*; 37.5% G + C) Cargo: *hnsT*, *hnsB*	PT4 GAL, NI: LT2
22	*SEN2085A–SEN2085D*	*SG2117–SG2120*	*ROD22*	4.9 kb	Group D LPS O-chain genes *rfbE* and *rfbS*	PT4 GAL. Different genes in LT2 (group B O-chain)
peg	*SEN2144A–SEN2145B*	*SG2182–SG2186*	*peg*	4.8 kb	Fimbrial operon	PT4 GAL, NI: LT2
SPI-17	*SEN2375A–SEN2380A*	*SG2425–SG2430A*	*SPI-17*	3.6 kb	Prophage remnant. Cargo: *gtrA*, *gtrB** *gtrC* (tRNA-*arg*)	PT4 GAL, NI: LT2 (Vernikos and Parkhill 2006)
25	*SEN2471–SEN2473*	*SG2522–SG2524*	*ROD25*	4 kb	Membrane transport system	PT4 GAL. Remnant in LT2 (STM2492*)
SPI-9	*SEN2609–SEN2612*	*SG2666–SG2671*	*SPI-9*	16.3 kb	*Salmonella* pathogenicity island (10Sa_RNA)	PT4 GAL LT2 (Parkhill et al. 2001)
SPI-1	*SEN2703–SEN2744*	*SG2764–SG2806*	*SPI-1*	40.2 kb	*Salmonella* pathogenicity island	PT4 GAL LT2 (Hansen-Wester and Hensel 2001)
28	*SEN2746–SEN2746A*	*SG2808–SG2811*	*ROD28*	2 kb	Membrane proteins	PT4 GAL, NI: LT2

(continued)

Table 2. *Continued*

Label[a]	PT4 CDS range	GAL CDS range	Locus name(s)[b]	Size in PT4[c]	General description of locus	Where present
ste	*SEN2794–SEN2799*	*SG2859–SG2864*	*ste*	5.6 kb	Fimbrial operon (*S.* Typhi E)	PT4 GAL. Remnant in LT2 (major pilin remnant: 3,102,016–3,102,150)
30	*SEN2864–SEN2878*	*SG2930**	*ROD30*	13 kb (0.16 kb)	*rcnA*, the *std* fimbrial, operon	PT4 LT2. Remnant in GAL
std	*SEN2871–SEN2873*	*NP*	*std*	5.0 kb	Within *ROD30*. Fimbrial operon (*S.* Typhi D)	PT4 LT2, NI: GAL
SPI-13	*SEN2960–SEN2966*	*SG3011–SG3017*	*SPI-13*	7.4 kb	*Salmonella* pathogenicity island (tRNA-*phe*)	PT4 GAL LT2 (Shah et al. 2005)
lpf	*SEN3459–SEN3463*	*SG3793–SG3798*	*lpf*	5.5 kb	Fimbrial operon (long polar fimbriae)	PT4 GAL LT2
SPI-3	*SEN3572–SEN3586*	*SG3665–SG3680*	*SPI-3*	16.6 kb	*Salmonella* pathogenicity island (tRNA-*selC*)	PT4 GAL LT2 (Blanc-Potard et al. 1999)
34	*SEN3896–SEN3898*	*SG3312–SG3314*	*ROD34*	4.2 kb	Amino acid metabolic CDS	PT4 GAL, NI: LT2
35	*SEN3978–SEN3981*	*SG4054–SG4058*	*ROD35*	4.5 kb	Unknown	PT4 GAL, NI: LT2
SPI-4	*SEN4026–SEN4032*	*SG4100–SG4105*	*SPI-4*	25 kb	*Salmonella* pathogenicity island	PT4 GAL LT2 (Wong et al. 1998; Parkhill et al. 2001)
37	*SEN4165–SEN4166*	*SG4241–SG4242*	*ROD37*	3 kb	Unknown	PT4 GAL, NI: LT2
SPI-10	*SEN4244–SEN4254*	*SG4311–SG4325*	*SPI-10*	10 kb	*Salmonella* pathogenicity island (tRNA-*leu*)	GAL PT4, NI: LT2 (Parkhill et al. 2001)
sef	*SEN4247–SEN4251*	*SG4318–SG4322*	*sef*	5.1 kb	Fimbrial operon (*Salmonella* Enteritidis fimbriae)	PT4 GAL, NI: LT2
40	*SEN4283–SEN4292*	*SG4349–SG4357*	*ROD40*	13.5 kb (10 kb)	Type I restriction-modification system	PT4. Degenerate in GAL, NI: LT2
sth	*SEN4347–SEN4351*	*SG4413–SG4417*	*sth*	5.5 kb	Fimbrial operon (*S.* Typhi H)	PT4 GAL LT2
42	*SEN3843*–SEN3844**	*SG3367–SC3368*	*ROD42*	1 kb	C4-dicarboxylate transporters	GAL LT2. Deleted from PT4

[a]Labels used to mark these regions on the outer ring of Figures 2 and 5.
[b]Square brackets indicate the name in *S.* Gallinarum 287/91.
[c]Numbers in parentheses indicate the size of the analogous region in *S.* Gallinarum. Only shown when significantly different from that found in *S.* Enteritidis PT4.
(*) Gene remnant; (DR) direct repeat; (NI) not detected in; (LPS) lipopolysaccharide locus; (LT2) *S.* Typhimurium LT2; (GAL) *S.* Gallinarum 287/91; (PT4) *S.* Enteritidis PT4; (Nal^r) nalidixic acid resistance; (Fos^r) fosfomycin resistance; (Det^r) detergent resistance.

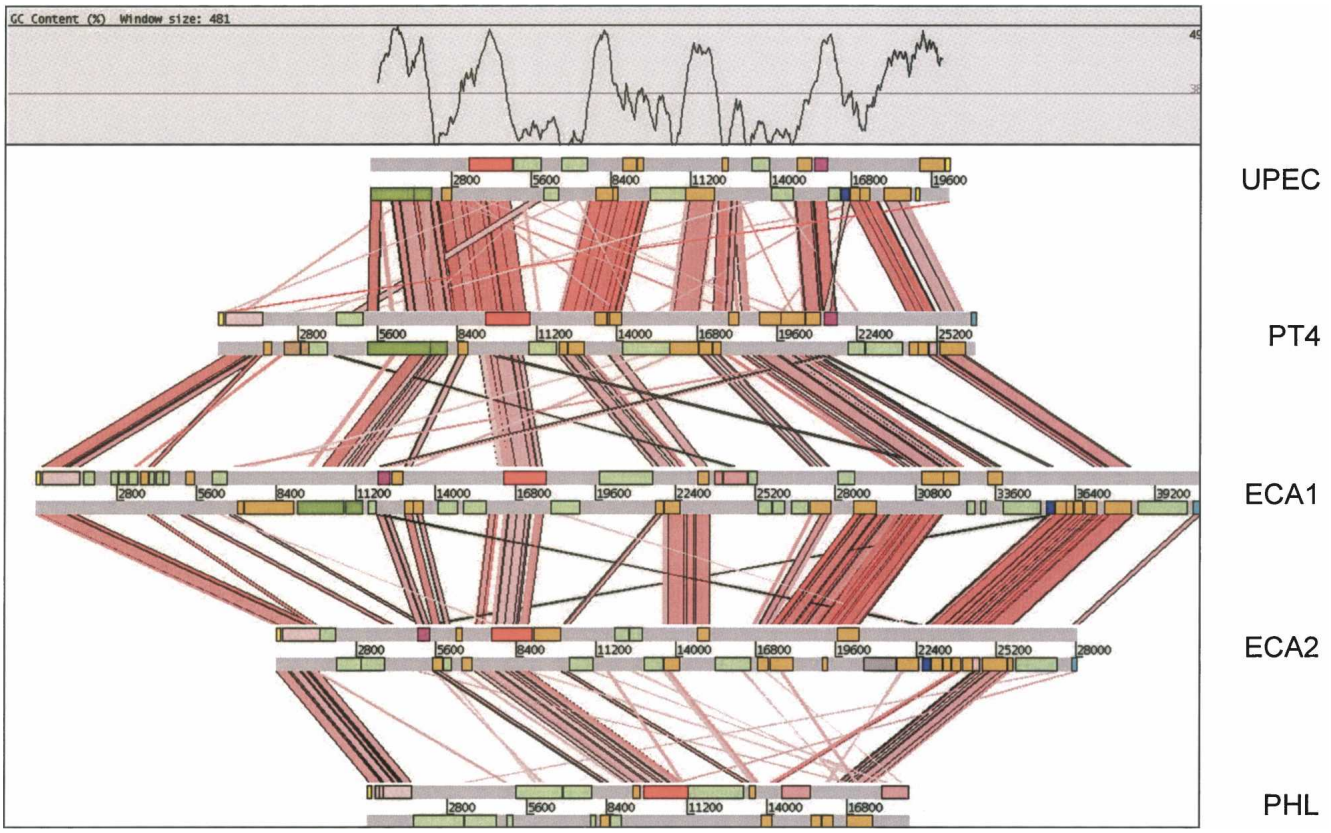


Figure 4. Comparison of the *ROD21* locus of *S.* Enteritidis with related genomic islands. ACT comparison (http://www.sanger.ac.uk/Software/ACT) of amino acid matches between the complete six-frame translations (computed using TBLASTX) of *ROD21* compared with related loci uropathogenic *E. coli* strain CFT073 (UPEC), *Erwinia carotovora* sbsp. *atroseptica* strain SCRI1043 (two loci: *ECA1* and *ECA2*), and *Photorhabidus luminescens*,sbsp. *laumondii* TT01 (PHL; see Methods). The red bars spanning between the genomes represent individual TBLASTX matches. CDS are marked as colored boxes positioned on the horizontal gray DNA bars: (orange) genes conserved in two or more of the genomic islands; (light green) variable genes of unknown function; (dark pink) *hnsB*; (dark blue) *hnsT*; (light pink) integrase; (dark green) type IV pilin-associated genes; (red) plasmid-related mobility functions; (salmon pink) transposase-related genes; (yellow) tRNA genes; (light blue)repeats. The G + C profile for the UPEC loci is shown above. The scale is marked in base pairs.

larity with *S.* Enteritidis PT4 compared with *S.* Typhimurium LT2 (Figs. 3, 5). The average nucleotide identities of orthologs shared between *S.* Gallinarum 287/91 and *S.* Enteritidis PT4 were higher (99.7%) than those found in LT2 (98.93%). Another obvious feature is the massive accumulation of pseudogenes in *S.* Gallinarum 287/91 compared with *S.* Enteritidis PT4 and *S.* Typhimurium LT2. The genome of *S.* Gallinarum 287/91 is slightly smaller than *S.* Enteritidis PT4, carries significantly fewer tRNA genes (Table 1), and is colinear except for a single inversion (817 kb; about the rRNA operons) and translocation of a region (49 kb) located between two different rRNA operons (Figs. 1, 5).

The number of CDS unique to *S.* Gallinarum 287/91 (76) or shared exclusively between *S.* Gallinarum 287/91 and *S.* Typhimurium LT2 (16) was small and predominantly phage-associated (Figs. 3, 5). Moreover, genes from both of these categories all fell within regions (*SPI-6*, *ROD9*, *ROD42*, and ϕSG12) that are present in *S.* Enteritidis PT4 but appear to be in the process of being lost. Consequently, these genes are unlikely to be recent acquisitions by *S.* Gallinarum 287/91.

Of the 130 CDS specific to *S.* Enteritidis PT4, compared with *S.* Gallinarum 287/91, those associated with *ROD4* and prophages ϕSE10, ϕSE14, and ϕSE20 (82 CDS) appear to be recent acquisitions with no evidence of them ever being present in *S.* Gallinarum 287/91(Fig. 3; Table 2). Of the remaining 48 CDS in this category, 21 were located on loci (*ROD15*, *ROD18*, and *ROD30*) (Table 2) almost entirely deleted from *S.* Gallinarum 287/91. The others were located on shared loci such as *ROD14* and *SPI-6* that are degenerate in both serotypes, compared with *S.* Typhimurium LT2. The functions that these RODs encode in *S.* Enteritidis PT4 are summarized in Table 2.

Thus, we provide compelling genetic evidence that *S.* Enteritidis and *S.* Gallinarum are recently diverged clones. On this conclusion, we have plotted the most parsimonious explanation for the observed gene flux following the divergence of *S.* Typhimurium LT2, *S.* Enteritidis PT4, and *S.* Gallinarum 287/91 (Fig. 6).

Functional gene loss and pseudogene formation

In addition to the large scale deletion, there is further evidence of reductive evolution in *S.* Gallinarum 287/91 in the form of 309 putative pseudogenes that carry frameshifts or premature stop codons or that are remnants of genes present in other bacteria. Remarkably, this represents ~7% of the total coding capacity of the genome and includes genes from many functional categories, including metabolism and virulence (for a full list, see Supplemental Table 1).

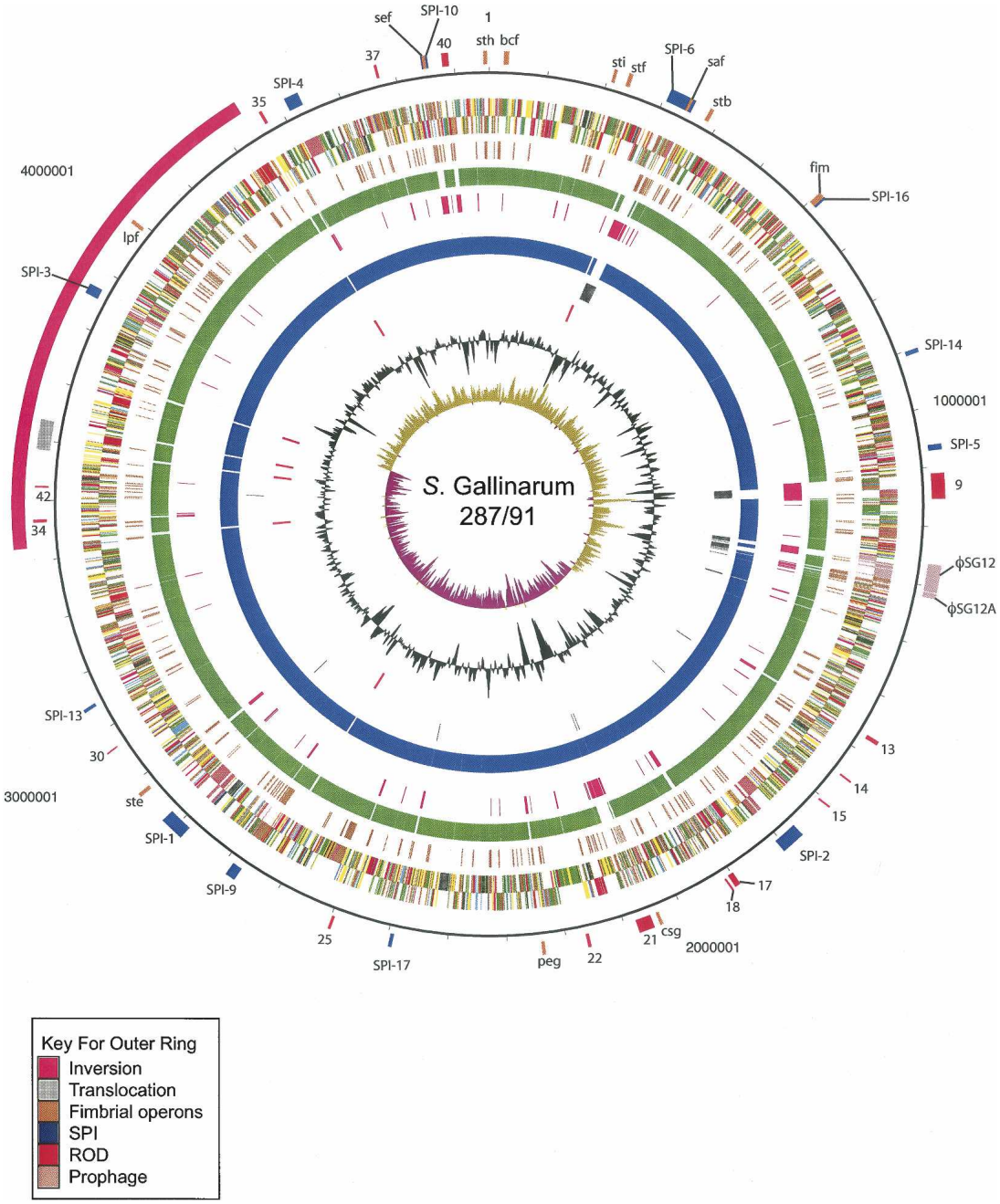


Figure 5. Circular representation of the *S.* Gallinarum 287/91 chromosome. For a full description of this figure, see legend to Figure 2. The exception is that circle 5 shows the position of *S.* Gallinarum 287/91 pseudogenes. Circles 6 and 8 show the position of *S.* Gallinarum 287/91 genes which have orthologs (by reciprocal FASTA analysis) in *S.* Typhimurium LT2 (all CDS colored green) or *S.* Enteritidis PT4 (all CDS colored blue), respectively. Circles 7 and 9 show the position of *S.* Gallinarum 287/91 genes that lack orthologs in (by reciprocal FASTA analysis) in *S.* Typhimurium strain LT2 (all CDS colored pink) and *S.* Enteritidis PT4 (all CDS colored gray), respectively. Circle 10 shows the position of the *S.* Gallinarum 287/91 rRNA operons (red).

S. Gallinarum 287/91 has lost several pathways, which are likely to narrow the spectrum of substrates available for use as sources of carbon and energy. These include *malS* encoding a periplasmic alpha-amylase required for growth on long chain maltodextrins (Schneider et al. 1992), those affecting the catabolism of (*D*)-glucarate (*gudD*) and the mutation in the *hyaF* gene (within the hydrogenase 1 [*hya*] gene cluster).

S. Gallinarum 287/91 also possesses multiple mutations in genes within all three operons required for the breakdown 1,2-propanediol: *ttr*, *cbi*, and *pdu* operons directing tetrathionate respiration; coenzyme B_{12} biosynthesis (B_{12}; cobalamine); and 1,2-propanediol degradation, respectively (Supplemental Table 1) (Roth et al. 1996). 1,2-Propanediol is an important source of energy for *S.* Typhimurium, and *cbi* mutants are significantly attenuated in their ability to grow in macrophages (Klumpp and Fuchs 2007). Consequently, for most of the salmonellae, the ability to degrade propanediol is the likely selective pressure maintaining the *cbi* and *ttr* genes, and the loss of function of any of

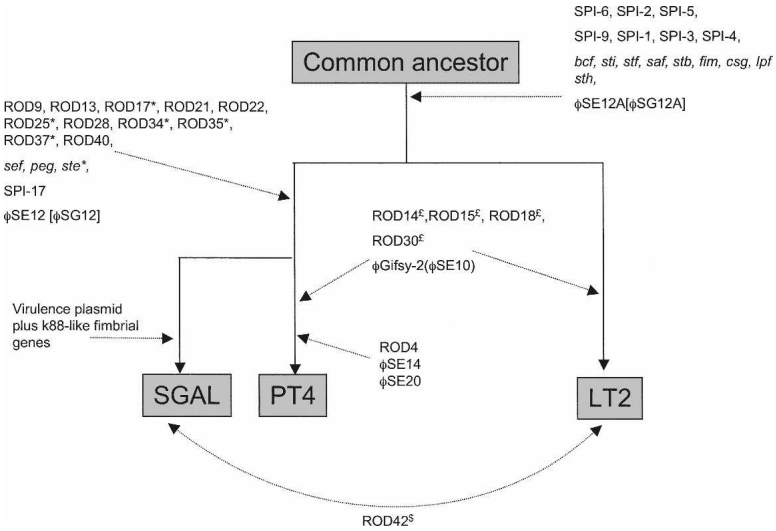


Figure 6. Line diagram to represent the whole-genome differences of *S.* Enteritidis PT4, *S.* Typhimurium strain LT2, and *S.* Gallinarum strain 287/91. A summary of the observed loss and gain of RODs described in Table 2. The diagram is based on the assumption that following the divergence of PT4 and LT2 from a common ancestor PT4 and GAL have subsequently diverged. Branches are not intended to infer phylogenetic distance. Evidence that a locus was once present in LT2, PT4, or GAL (see legend to Figure 1) but has subsequently been deleted from that genome is marked by the suffix *, $, or £, respectively. Brackets indicate the name for locus in GAL. Parentheses indicate the name for locus in PT4. Dotted arrows mark the position in the pseudo-tree at which that ROD(s) appears

these three pathways was probably the precursor to the degeneration of the others in *S.* Gallinarum 287/91. Mutations within the *cbi*, *pdu*, and *ttr* genes was also a feature of the *S.* Typhi CT18 genome (Table 3), suggesting that these mutations may be characteristic of more invasive *Salmonella* serotypes.

Uniquely among the *Salmonella*, *S.* Gallinarum 287/91 and *S.* Pullorum are unable to make glycogen (Supplemental Table 2) (McMeechan et al. 2005). The genome sequence data revealed that mutations in the glycogen biosynthetic pathway are extensive in *S.* Gallinarum 287/91, including *glgA*, *glgB*, and *glgC*, which are responsible for all steps in biosynthesis. Although the significance of these mutations are not clear, they may explain in part the poor survival of this bacterium outside of the host (McMeechan et al. 2005).

Several of the identified *S.* Gallinarum 287/91 pseudogenes lie in amino acid catabolic or biosynthetic pathways. *Salmonella* encodes three pathways for arginine degradation (Reitzer 2003), and *S.* Gallinarum 287/91 carries mutations in two of these: There are deletions or frameshifts in the genes of the arginine *N*-succinyltransferase (AST) (*astA*) and arginine deiminase (ADI) (*arcA*) pathways, respectively. *S.* Gallinarum 287/91 also carries a mutation in *speC* encoding the ornithine decarboxylase, making the one remaining intact arginine catabolic pathway, arginine decarboxylase pathway (ADC), an essential biosynthetic route for putrescine. The mutation in *speC* is also likely to explain the inability of *S.* Gallinarum to decarboxylate ornithine, a defining feature of this *Salmonella* serovar (Supplemental Table 2) (Crichton and Old 1990).

Pseudogenes potentially involved in virulence and host adaptation

S. Gallinarum 287/91 is defined as being nonmotile. *S.* Gallinarum 287/91 carries 50 genes associated with motility and chemotaxis, distributed over three loci. Of these, five genes, present in two loci, carry mutations that explain the nonmotile phenotype, including *cheM*, *flhA*, *flhB*, *flgK*, and *flgI* (Supplemental Table 1).

Of the 13 fimbrial operons detected in *S.* Enteritidis PT4, the *std* fimbrial operon is not present in *S.* Gallinarum 287/91 (see *ROD30* above). The remaining 12 *S.* Gallinarum 287/91 fimbrial operons are identical to those in *S.* Enteritidis PT4 except for the mutations in genes within operons *sti*, *stf*, *saf*, *stb*, *peg*, *lpf*, *sef*, and *sth* (see Supplemental Table 1). The level of pseudogene formation within these 12 fimbrial gene clusters (16%) is over twice that of the genome average (7%). Only operons *fim*, *bcf*, *csg*, and *ste* remain undisrupted on the *S.* Gallinarum 287/91 chromosome (summarized in Fig. 1). However, both *S.* Enteritidis and *S.* Gallinarum carry fimbrial operons on their virulence plasmids. The *S.* Enteritidis virulence plasmid carries five genes—*pefA*, *pefB*, *pefC*, *pefD*, and *pefR*—highly conserved with those of *S.* Typhimurium LT2 (Woodward and Kirwan 1996). The *pef* operon is not present on the *S.* Gallinarum 287/91 plasmid; in its place are three intact fimbrial genes displaying sequence similarity with those of the *E. coli* K88 fimbrial gene cluster (Rychlik et al. 1998). This fimbrial operon represents the only evidence of *S.* Gallinarum 287/91 having acquired new functions since the split from *S.* Enteritidis. It would be interesting to know if any isolates of *S.* Enteritidis carry such a fimbrial operon and whether this is a common characteristic of all *S.* Gallinarum strains. Significantly, the host-adapted *S.* Typhi harbors novel fimbrial genes, including Type IV pili associated with the acquisition of mobile elements (Pickard et al. 2003), and like *S.* Gallinarum 287/91, there is an elevated level of mutation in fimbrial genes (14% compared with the genome average of 4.4%), again suggesting parallel paths toward host adaptation (Table 3).

Salmonella can express several paralogous TTSS effector proteins, which show a degree of functional redundancy, for example, *sopE* and *sopE2* (Friebel et al. 2001). Other effectors related by sequence include *pipB* and *pipB2*, and *sifA* and *sifB*. *S.* Gallinarum 287/91 has lost one of each of these parlogous pairs. Other *S.* Gallinarum 287/91 TTSS effector genes that carry mutations include *sopA*, which has been implicated in *S.* Typhimurium–induced intestinal inflammation (Zhang et al. 2006). Using an antibody against the C-terminal portion of SopA, we detected a secreted protein of the expected size in *S.* Enteritidis PT4 but not *S.* Gallinarum 287/91 (data not shown), consistent with the location of a stop codon prior to the mAb-binding region in *S.* Gallinarum SopA predicted by the genome sequence (Supplemental Table 2). SopA influences *Salmonella*-induced enteritis, and taken together with the attrition of other Type III secreted effectors, this may partially dictate the differential virulence of the serovars in mammalian hosts.

As well as additional pseudogenes associated with cell interactions, again like *S.* Typhi, *S.* Gallinarum 287/91 carries mutations in genes also associated with shedding (*shdA* and *ratB*), drug resistance, DNA restriction/modification, and protective re-

Table 3. Summary of the common traits identified among the functions of genes lost independently by *S.* Typhi CT18 and *S.* Gallinarum 287/91

Process/pathway	*S.* Gallinarum 278/91[a]	*S.* Typhi CT18[b]
Cell interactions	*slrP, sopA, sifB, sspH2, sinH, pipB2, pagK, bigA*	*sopD2 (STY0971), sopA (STY2275), sopE2 (STY1987), sseJ (STY1439a), cigR (STY4024), misL (STY4030), marT (STY4027), sivH (STY2767), slrP (STY0833), bigA (STY4318)*
Fecal shedding	*shdA, ratB*	*shdA (STY2755), ratB (STY2758), sivH (STY2767)*
Fimbriae	*stdΔ, stiC, stfF, safC, stbC, pegC, lpfC, sefD, sefC, sthB, sthA, sthE*	*bcfC (STY0026), fimI (STY0590), steA (STY3084), safE (STY0333), stgC (STY3920), ushA (STY0539), sefA (STY4836a) sefD (STY4839), sefR (STY4841), sthC (STY4938), sthE (STY4938)*
Flagella/motility	*cheM, flhB, flhA, flgK, flgI*	*fliB (STY2166)*
Type I restriction modification	*hsdR, hsdM*	*hsdM (STY4833)*
Type III restriction modification		
Restriction enzyme *Sty*LT1	*mod*	*res (STY0389)*
Cobalamine biosynthesis	*pocR, cobD, cbiD, cbiC, cbiO*	*cbiM (STY2226), cbiK (STY2229), cbiJ (STY2231)*
Propanediol utilization	*pduG, pduO*	*pduN (STY2254)*
Metal/drug resistance and transport		
Copper	*cusA, cusS*	*cusA (STY0610), cusS (STY0609a)*
Nickel/cobalt	*rcnAcusA, cusS*	*rcnA (STY3169)*
Nickel	*nxiArcnA*	*nxiA (STY2901)*
Acriflavin	*acrFnxiA*	*acrE (STY3569), acrF (STY3570)*
Tetrathionate respiration	*ttrB, ttrC*	*ttrS (STY1735)*
Trehalose degradation/synthesis	*treC*	*treA (STY1924)*
Hydrogenase I	*hyaF*	*hyaB2 (STY1525), hyaA (STY1319)*
Ornithine catabolism	*speC*	*speC (STY3270), speF (STY0739)*
Amino acid catabolism		
L-serine/*L*-threonine	*tdcG*	*tdcC (STY3426)*
Cellulose biosynthesis	*bcsG*	*bcsC (STY4184)*
Surface polysaccharide		
LPS O-chain	*gtrB*	*gtrB (STY2627b)*
LPS core	*rfaZ* (*waaZ*)	*wcaK(STY2311), wcaD (STY2324), wcaA (STY2328)*
Alternative terminal electron acceptors		
Dimethyl sulfoxide reductase	*dmsA2, dmsA1*	*dmsA2 (STY4503), dmsB2 (STY4506)*
Trimethylamine *N*-oxide (TMAO)	*torS*	*torR (STY3954), torC (STY3955)*
Carbon source		
D-Glucarate uptake and degradation	*gudD*	*gudP (STY4097)*
Maltodextrin and Maltose associated	*malS, malY, malX*	*malY (STY1657a), malX (STY1657)*

[a]For systematic gene identifiers and a description of function, see Supplemental Table 1.
[b]Parkhill et al. 2001.

sponses (Supplemental Table 1). The majority of *S.* Enteritidis isolates can produce a biofilm, of which cellulose is a key component. While mutations in biofilm production may not measurably affect virulence, they are significantly less resistant to chemical and mechanical stress. Consequently, this is likely to be an adaptation by *Salmonella* to survival in the environment but has also been suggested to prolong retention in the gut (Solano et al. 2002). We have shown experimentally that *S.* Gallinarum 287/91 is unable to make cellulose, and this is likely to be explained by a mutation in *bcsG* (Supplemental Table 2). This is consistent with the reduced ability of *S.* Gallinarum to colonize the gut compared with *S.* Enteritidis.

Conclusions

The data presented in this report provide several clear messages, some of which may be experimentally tractable. Comparative analysis of the genomes of *S.* Enteritidis PT4 and *S.* Gallinarum 287/91 shows that representative strains of these two *S. enterica* serovars are highly related and that *S.* Gallinarum may be a direct descendant of *S.* Enteritidis. Importantly, *S.* Enteritidis is promiscuous, being able to colonize and infect multiple hosts, including chickens, cattle, mice, and humans, in addition to producing murine typhoid. Whereas *S.* Gallinarum is highly restricted to causing a typhoid-like disease in avian species, it is relatively noninfectious in other hosts, including mice, and does not colonize the gut of animals. Thus, we suggest that there is an experimental opportunity to use genetic approaches to define the genetic basis of host restriction by directly comparing the pathogenicity of strains of *S.* Enteritidis and *S.* Gallinarum in murine and chicken models.

Previous genome analyses on host-restricted salmonellae has involved human-restricted serovars, including *S.* Typhi and *S.* Paratyphi, limiting experimental tractability. Nevertheless genome comparisons of host-restricted/adapted *S. enterica* serovars, and indeed of other pathogens, indicate that loss of gene function may be a common evolutionary mechanism through which host adaptation occurs. Gene loss not only may limit the interhost promiscuity of the pathogen but also is likely to restrict the potential pathogenicity in the host to a more limited set of interactions. We hypothesize that gene loss may be a mechanism of targeting the invading pathogen preferentially to particular tissues or host cells and avoiding the potential stimulation of nonspecific inflammation. An example here would be the loss of flagella or fimbriae, which can mediate attachment and invasion of cell surfaces and may activate pattern recognition molecules.

In addition, gene loss can influence the ability of the pathogen to survive in the external environment or even in stressful situations within the host. Table 3 provides a list of some of the common traits identified among the functions of genes lost independently by *S.* Typhi and *S.* Gallinarum. Some of the overlaps are striking, including the loss of common TTSS effectors and genes involved in common metabolic processes such as cobalamine and propanediol utilization, tetrathionate respiration, sugar uptake and utilization, hydrogenase activity, cellulose production, ornithine decarboxylase activity, and electron transport acceptor function. Some of these common traits have also been noted to have changed in representatives of gut adapted (*Y. enterocolitica*) versus systemic (*Y. pestis)* yersiniae, and again in this system, gene loss may be involved in the adaptation from a gut to a systemic lifestyle. We believe that further studies analyzing the contribution of pseudogenes and their functional alleles to host adaptation and tissue specificity and, in particular, the parallel but overlapping degradative evolutionary pathways followed by different organisms adapting to different hosts will lead to significant understanding of the mechanisms of host adaptation and host restriction and could be applicable to the less tractable human-adapted organisms, such as *S.* Typhi.

Methods

Bacterial strains

S. Gallinarum strain 287/91 was isolated from an outbreak of fowl typhoid in brown egg–laying hens by A. Berchieri, University of Sao Paulo, Jaboticabal, Brazil. It is highly virulent (>90% mortality) in susceptible breeds of chickens (P. Barrow and A. Berchier, unpubl.). It was chosen in preference to the well-characterized strain 9 (Smith 1955) because of the length of laboratory passage of the latter strain. *S.* Enteritidis phage type 4 (PT4) strain P125109 was isolated from an outbreak of human food-poisoning in the United Kingdom that was traced back to a poultry farm. The strain is highly virulent in newly hatched chickens and is also invasive in laying hens, resulting in egg contamination (Barrow 1991; Barrow and Lovell 1991). Biochemical tests for carbohydrate catabolism were performed using api 50 CH according to manufacturer's instructions (BioMerieux).

Growth and sequencing of *S.* Enteritidis PT4 and *S.* Gallinarum 287/91

Methods for sequencing *S.* Enteritidis PT4 and *S.* Gallinarum 287/91 were identical unless stated. A single bacterial colony was picked from Congo Red agar and grown overnight in BAB broth with shaking at 37°C. Cells were collected, and total DNA (10 mg) was isolated using proteinase K treatment followed by phenol extraction. The DNA was fragmented by sonication, and several libraries were generated in pUC18 using size fractions ranging from 1.0–2.5 kb.

The whole genome sequenced to a depth of 9× coverage from M13mp18 (insert size 1.4–2 kb) and pUC18 (insert size 2.2–4.2 kb) small insert libraries using dye terminator chemistry on ABI3700 automated sequencers. End sequences from larger insert plasmid (pBACe3.6, 12–30 kb insert size) libraries were used as a scaffold.

The sequence was assembled, finished, and annotated as described previously (Parkhill et al. 2000), using the program Artemis (Berriman and Rutherford 2003) to collate data and facilitate annotation.

The genomes have been submitted to EMBL under the following accession numbers: *S.* Enteritidis PT4 genome, AM933172; *S.* Gallinarum 287/91 genome, AM933173.

In silico genome analysis

The genome sequences of *S.* Typhimurium strain LT2 (McClelland et al. 2001), *S.* Enteritidis PT4, and *S.* Gallinarum 287/91 were compared pairwise using the Artemis Comparison Tool (ACT) (Carver et al. 2005). Subsequences taken from the genomes of uropathogenic *E. coli* strain CFT073 (Welch et al. 2002), *Erwinia carotovora* sbsp. *atroseptica* strain SCRI1043 (Bell et al. 2004), and *Photorhabidus luminescens*,sbsp. *laumondii* TT01 (Duchaud et al. 2003) were compared with ACT as above and used to construct Figure 4.

Pseudogenes had one or more mutations that would ablate expression; each of the inactivating mutations was confirmed by subsequently rechecking the original sequencing data and where necessary were resequenced.

Orthologous gene sets were identified by reciprocal FASTA searches. Only those pairs of homologous CDS were retained for further analysis where the predicted amino acid identity was ≥40% over 80% of the protein length. These genes were then subject to manual curation using gene synteny to increase the accuracy of this analysis. This strategy was applied to pairwise comparisons of the genomes of *S.* Typhimurium strain LT2, *S.* Enteritidis PT4, and *S.* Gallinarum 287/91.

Cellulose production assay

For preparation and use of Calcofluor plates, Calcofluor white stain was obtained from Sigma as a 0.1% w/v solution. This was added to L-agar at a final concentration of 200 µg/mL as recommended by Solano et al. (2002). Bacterial cultures were inoculated and then left at room temperature for 48 h.

Colony fluorescence was examined by holding the plate over a 366-nm UV transilluminator. Controls used included *E. coli* C600 (negative control) and *S. typhimurium* SL1344 (positive control). Colony fluorescence was scored quantitatively using the controls as standards.

Acknowledgments

We thank the core sequencing and informatics teams at the Sanger Institute for their assistance and The Wellcome Trust for its support of the Sanger Institute Pathogen Sequencing Unit. This project was funded through The Wellcome Trust Beowulf Genomics Initiative.

References

Barrow, P. 1991. Experimental infection of chickens with *Salmonella enteritidis*. *Avian Pathol.* **20:** 145–153.

Barrow, P.A. 2000. The paratyphoid salmonellae. *Rev. Sci. Tech.* **19:** 351–375.

Barrow, P.A. and Lovell, M.A. 1991. Experimental infection of egg-laying hens with *Salmonella enteritidis*. *Avian Pathol.* **20:** 339–352.

Bausch, C., Peekhaus, N., Utz, C., Blais, T., Murray, E., Lowary, T., and Conway, T. 1998. Sequence analysis of the GntII (subsidiary) system for gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in *Escherichia coli*. *J. Bacteriol.* **180:** 3704–3710.

Bell, K.S., Sebaihia, M., Pritchard, L., Holden, M.T., Hyman, L.J., Holeva, M.C., Thomson, N.R., Bentley, S.D., Churcher, L.J., Mungall, K., et al. 2004. Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. *Proc. Natl. Acad. Sci.* **101:** 11105–11110.

Berriman, M. and Rutherford, K. 2003. Viewing and annotating sequence data with Artemis. *Brief. Bioinform.* **4:** 124–132.

Bishop, A.L., Baker, S., Jenks, S., Fookes, M., Gaora, P.O., Pickard, D., Anjum, M., Farrar, J., Hien, T.T., Ivens, A., et al. 2005. Analysis of

the hypervariable region of the *Salmonella enterica* genome associated with tRNA(*leuX*). *J. Bacteriol.* **187:** 2469–2482.
Blanc-Potard, A.B., Solomon, F., Kayser, J., and Groisman, E.A. 1999. The SPI-3 pathogenicity island of *Salmonella enterica*. *J. Bacteriol.* **181:** 998–1004.
Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., and Parkhill, J. 2005. ACT: The Artemis Comparison Tool. *Bioinformatics* **21:** 3422–3423.
Chiu, C.H., Tang, P., Chu, C., Hu, S., Bao, Q., Yu, J., Chou, Y.Y., Wang, H.S., and Lee, Y.S. 2005. The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. *Nucleic Acids Res.* **33:** 1690–1698.
Collighan, R.J. and Woodward, M.J. 2001. The SEF14 fimbrial antigen of *Salmonella enterica* serovar Enteritidis is encoded within a pathogenicity islet. *Vet. Microbiol.* **80:** 235–245.
Cooke, F.J., Wain, J., Fookes, M., Ivens, A., Thomson, N., Brown, D.J., Threlfall, E.J., Gunn, G., Foster, G., and Dougan, G. 2007. Prophage sequences defining hot spots of genome variation in *Salmonella enterica* serovar Typhimurium can be used to discriminate between field isolates. *J. Clin. Microbiol.* **45:** 2590–2598.
Crichton, P.B. and Old, D.C. 1990. Salmonellae of serotypes *gallinarum* and *pullorum* grouped by biotyping and fimbrial-gene probing. *J. Med. Microbiol.* **32:** 145–152.
Doyle, M., Fookes, M., Ivens, A., Mangan, M.W., Wain, J., and Dorman, C.J. 2007. An H-NS-like stealth protein aids horizontal DNA transmission in bacteria. *Science* **315:** 251–252.
Duchaud, E., Rusniok, C., Frangeul, L., Buchrieser, C., Givaudan, A., Taourit, S., Bocs, S., Boursaux-Eude, C., Chandler, M., Charles, J.F., et al. 2003. The genome sequence of the entomopathogenic bacterium *Photorhabdus luminescens*. *Nat. Biotechnol.* **21:** 1307–1313.
Edwards, R.A., Schifferli, D.M., and Maloy, S.R. 2000. A role for *Salmonella* fimbriae in intraperitoneal infections. *Proc. Natl. Acad. Sci.* **97:** 1258–1262.
Eriksson, S., Lucchini, S., Thompson, A., Rhen, M., and Hinton, J.C. 2003. Unravelling the biology of macrophage infection by gene expression profiling of intracellular *Salmonella enterica*. *Mol. Microbiol.* **47:** 103–118.
Figueroa-Bossi, N., Uzzau, S., Maloriol, D., and Bossi, L. 2001. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. *Mol. Microbiol.* **39:** 260–271.
Friebel, A., Ilchmann, H., Aepfelbacher, M., Ehrbar, K., Machleidt, W., and Hardt, W.D. 2001. SopE and SopE2 from *Salmonella typhimurium* activate different sets of RhoGTPases of the host cell. *J. Biol. Chem.* **276:** 34035–34040.
Hansen-Wester, I. and Hensel, M. 2001. Salmonella pathogenicity islands encoding type III secretion systems. *Microbes Infect.* **3:** 549–559.
Hensel, M. 2000. *Salmonella* pathogenicity island 2. *Mol. Microbiol.* **36:** 1015–1023.
Iqbal, M., Philbin, V.J., Withanage, G.S., Wigley, P., Beal, R.K., Goodchild, M.J., Barrow, P., McConnell, I., Maskell, D.J., Young, J., et al. 2005. Identification and functional characterization of chicken toll-like receptor 5 reveals a fundamental role in the biology of infection with *Salmonella enterica* serovar typhimurium. *Infect. Immun.* **73:** 2344–2350.
Kaiser, P., Rothwell, L., Galyov, E.E., Barrow, P.A., Burnside, J., and Wigley, P. 2000. Differential cytokine expression in avian cells in response to invasion by *Salmonella typhimurium*, *Salmonella enteritidis* and *Salmonella gallinarum*. *Microbiology* **146:** 3217–3226.
Kilger, G. and Grimont, P.A. 1993. Differentiation of *Salmonella* phase 1 flagellar antigen types by restriction of the amplified *fliC* gene. *J. Clin. Microbiol.* **31:** 1108–1110.
Klumpp, J. and Fuchs, T.M. 2007. Identification of novel genes in genomic islands that contribute to *Salmonella typhimurium* replication in macrophages. *Microbiology* **153:** 1207–1220.
Li, J., Smith, N.H., Nelson, K., Crichton, P.B., Old, D.C., Whittam, T.S., and Selander, R.K. 1993. Evolutionary origin and radiation of the avian-adapted non-motile salmonellae. *J. Med. Microbiol.* **38:** 129–139.
McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latreille, P., Courtney, L., Porwollik, S., Ali, J., Dante, M., Du, F., et al. 2001. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* **413:** 852–856.
McClelland, M., Sanderson, K.E., Clifton, S.W., Latreille, P., Porwollik, S., Sabo, A., Meyer, R., Bieri, T., Ozersky, P., McLellan, M., et al. 2004. Comparison of genome degradation in Paratyphi A and Typhi: Human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nat. Genet.* **36:** 1268–1274.
McMeechan, A., Lovell, M.A., Cogan, T.A., Marston, K.L., Humphrey, T.J., and Barrow, P.A. 2005. Glycogen production by different *Salmonella enterica* serotypes: Contribution of functional *glgC* to virulence, intestinal colonization and environmental survival. *Microbiology* **151:** 3969–3977.
Mmolawa, P.T., Schmieger, H., Tucker, C.P., and Heuzenroeder, M.W. 2003. Genomic structure of the *Salmonella enterica* serovar Typhimurium DT 64 bacteriophage ST64T: Evidence for modular genetic architecture. *J. Bacteriol.* **185:** 3473–3475.
Navarre, W.W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S.J., and Fang, F.C. 2006. Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*. *Science* **313:** 236–238.
Nishino, K. and Yamaguchi, A. 2001. Analysis of a complete library of putative drug transporter genes in *Escherichia coli*. *J. Bacteriol.* **183:** 5803–5812.
Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S., et al. 2000. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403:** 665–668.
Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T., et al. 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413:** 848–852.
Pickard, D., Wain, J., Baker, S., Line, A., Chohan, S., Fookes, M., Barron, A., Gaora, P.O., Chabalgoity, J.A., Thanky, N., et al. 2003. Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7. *J. Bacteriol.* **185:** 5055–5065.
Raleigh, E.A. 1992. Organization and function of the *mcrBC* genes of *Escherichia coli* K-12. *Mol. Microbiol.* **6:** 1079–1086.
Reitzer, L. 2003. Nitrogen assimilation and global regulation in *Escherichia coli*. *Annu. Rev. Microbiol.* **57:** 155–176.
Rodrigue, D.C., Tauxe, R.V., and Rowe, B. 1990. International increase in *Salmonella enteritidis*: A new pandemic? *Epidemiol. Infect.* **105:** 21–27.
Roth, J.R., Lawrence, J.G., and Bobik, T.A. 1996. Cobalamin (coenzyme B12): synthesis and biological significance. *Annu. Rev. Microbiol.* **50:** 137–181.
Rychlik, I., Lovell, M.A., and Barrow, P.A. 1998. The presence of genes homologous to the K88 genes *faeH* and *faeI* on the virulence plasmid of *Salmonella gallinarum*. *FEMS Microbiol. Lett.* **159:** 255–260.
Schneider, E., Freundlieb, S., Tapio, S., and Boos, W. 1992. Molecular characterization of the MalT-dependent periplasmic alpha-amylase of *Escherichia coli* encoded by *malS*. *J. Biol. Chem.* **267:** 5148–5154.
Shah, D.H., Lee, M.J., Park, J.H., Lee, J.H., Eo, S.K., Kwon, J.T., and Chae, J.S. 2005. Identification of *Salmonella gallinarum* virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis. *Microbiology* **151:** 3957–3968.
Shivaprasad, H.L. 2000. Fowl typhoid and pullorum disease. *Rev. Sci. Tech.* **19:** 405–424.
Smith, H.W. 1955. Observations on experimental fowl typhoid. *J. Comp. Pathol.* **65:** 37–54.
Solano, C., Garcia, B., Valle, J., Berasain, C., Ghigo, J.M., Gamazo, C., and Lasa, I. 2002. Genetic analysis of *Salmonella enteritidis* biofilm formation: Critical role of cellulose. *Mol. Microbiol.* **43:** 793–808.
Stanley, T.L., Ellermeier, C.D., and Slauch, J.M. 2000. Tissue-specific gene expression identifies a gene in the lysogenic phage Gifsy-1 that affects *Salmonella enterica* serovar Typhimurium survival in Peyer's patches. *J. Bacteriol.* **182:** 4406–4413.
Sweeney, N.J., Laux, D.C., and Cohen, P.S. 1996. *Escherichia coli* F-18 and *E. coli* K-12 *eda* mutants do not colonize the streptomycin-treated mouse large intestine. *Infect. Immun.* **64:** 3504–3511.
Thomson, N., Baker, S., Pickard, D., Fookes, M., Anjum, M., Hamlin, N., Wain, J., House, D., Bhutta, Z., Chan, K., et al. 2004. The role of prophage-like elements in the diversity of *Salmonella enterica* serovars. *J. Mol. Biol.* **339:** 279–300.
Thomson, N.R., Howard, S., Wren, B.W., Holden, M.T., Crossman, L., Challis, G.L., Churcher, C., Mungall, K., Brooks, K., Chillingworth, T., et al. 2006. The complete genome sequence and comparative genome analysis of the high pathogenicity *Yersinia enterocolitica* strain 8081. *PLoS Genet.* **2:** e206. doi: 10.1371/journal.pgen.0020206.
Titheradge, A.J., Ternent, D., and Murray, N.E. 1996. A third family of allelic hsd genes in *Salmonella enterica*: Sequence comparisons with related proteins identify conserved regions implicated in restriction of DNA. *Mol. Microbiol.* **22:** 437–447.
Townsend, S.M., Kramer, N.E., Edwards, R., Baker, S., Hamlin, N., Simmonds, M., Stevens, K., Maloy, S., Parkhill, J., Dougan, G., et al. 2001. *Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences. *Infect. Immun.* **69:** 2894–2901.
Vernikos, G.S. and Parkhill, J. 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: Revisiting the

Salmonella pathogenicity islands. *Bioinformatics* **22:** 2196–2203.
Welch, R.A., Burland, V., Plunkett III, G., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci.* **99:** 17020–17024.
Williamson, H.S. and Free, A. 2005. A truncated H-NS-like protein from enteropathogenic *Escherichia coli* acts as an H-NS antagonist. *Mol. Microbiol.* **55:** 808–827.
Wong, K.K., McClelland, M., Stillwell, L.C., Sisk, E.C., Thurston, S.J., and Saffer, J.D. 1998. Identification and sequence analysis of a 27-kilobase chromosomal fragment containing a *Salmonella* pathogenicity island located at 92 minutes on the chromosome map of *Salmonella enterica* serovar typhimurium LT2. *Infect. Immun.* **66:** 3365–3371.
Wood, M.W., Jones, M.A., Watson, P.R., Hedges, S., Wallis, T.S., and Galyov, E.E. 1998. Identification of a pathogenicity island required for *Salmonella* enteropathogenicity. *Mol. Microbiol.* **29:** 883–891.
Woodward, M.J. and Kirwan, S.E. 1996. Detection of *Salmonella enteritidis* in eggs by the polymerase chain reaction. *Vet. Rec.* **138:** 411–413.
Zhang, Y., Higashide, W.M., McCormick, B.A., Chen, J., and Zhou, D. 2006. The inflammation-associated *Salmonella* SopA is a HECT-like E3 ubiquitin ligase. *Mol. Microbiol.* **62:** 786–793.

Received February 12, 2008; accepted in revised form June 17, 2008.

Journal of Bacteriology

Genome Sequences of Salmonella enterica Serovar Typhimurium, Choleraesuis, Dublin, and Gallinarum Strains of Well-Defined Virulence in Food-Producing Animals

Emily J. Richardson, Bhakti Limaye, Harshal Inamdar, Avik Datta, K. Sunitha Manjari, Gillian D. Pullinger, Nicholas R. Thomson, Rajendra R. Joshi, Michael Watson and Mark P. Stevens

J. Bacteriol. **2011, 193(12):3162. DOI: 10.1128/JB.00394-11. Published Ahead of Print 8 April 2011.**

JOURNAL OF BACTERIOLOGY, June 2011, p. 3162–3163 Vol. 193, No. 12
0021-9193/11/$12.00 doi:10.1128/JB.00394-11

Genome Sequences of *Salmonella enterica* Serovar Typhimurium, Choleraesuis, Dublin, and Gallinarum Strains of Well-Defined Virulence in Food-Producing Animals[▽]

Emily J. Richardson,[1] Bhakti Limaye,[2] Harshal Inamdar,[2] Avik Datta,[2] K. Sunitha Manjari,[2] Gillian D. Pullinger,[3] Nicholas R. Thomson,[4] Rajendra R. Joshi,[2] Michael Watson,[1] and Mark P. Stevens[1]*

The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, United Kingdom[1]; Centre for Development of Advanced Computing, University of Pune Campus, Pune 411007, India[2]; Enteric Bacterial Pathogens Laboratory, Institute for Animal Health, Compton, Berkshire, RG20 7NN, United Kingdom[3]; and The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom[4]

Received 23 March 2011/Accepted 31 March 2011

Salmonella enterica **is an animal and zoonotic pathogen of worldwide importance and may be classified into serovars differing in virulence and host range. We sequenced and annotated the genomes of serovar Typhimurium, Choleraesuis, Dublin, and Gallinarum strains of defined virulence in each of three food-producing animal hosts. This provides valuable measures of intraserovar diversity and opportunities to formally link genotypes to phenotypes in target animals.**

Salmonella enterica causes salmonellosis in humans and other warm-blooded animals. Over 2,600 serovars have been classified according to the reactivity of antisera to somatic lipopolysaccharide and flagellar antigens and are broadly grouped on the basis of host range and disease presentation. The molecular basis of the differential virulence and tropism of serovars remains ill defined (20). An understanding of such processes is required to develop strategies for disease control and to predict the threat posed by isolates from animals.

The extent to which currently sequenced strains are typical of the wider serovar is open to question. We report the sequencing and annotation of four strains representing serovars that produce significant illness in food-producing animals: *S*. Typhimurium strain ST4/74 (11), originally isolated from a calf with salmonellosis in the United Kingdom (17) and the parent of the widely used mouse virulent *hisG* auxotroph SL1344 (10); *S*. Choleraesuis var. kunzendorf strain SCSA50, a field isolate from a case of swine typhoid in the United Kingdom (3); *S*. Dublin strain SD3246, a Vi-negative isolate from a calf with systemic salmonellosis in the United Kingdom (24); and *S*. Gallinarum SG9, first described to cause fowl typhoid in orally dosed chickens by Smith in 1955 (19). Crucially, the virulence of each strain has been reciprocally compared in calves, pigs, and chickens (3, 4, 6, 14, 15, 16, 17, 24, 25, 26, 27), fulfilling Koch's postulates and enabling strain genotypes to be linked to phenotypes in target hosts.

Sequencing and annotation. 36 cycle paired-end sequencing was carried out on an Illumina GAIIx, yielding between 80 and 150X coverage. SOAPdenovo (13) was used to generate *de novo* contigs, and reads aligned to a reference using Novoalign (Novocraft, Selangor, Malaysia). *S*. Typhimurium 4/74 reads were assembled on the genome and large plasmid of strain SL1344 (http://www.sanger.ac.uk/Projects/Salmonella/). *S*. Choleraesuis SCSA50 reads were assembled on the genome of strain SC-B67 (7) and its virulence plasmid (28). *S*. Dublin SD3246 reads were assembled on the genome of strain CT_02021853 (accession no. CP001144). *S*. Gallinarum SG9 reads were assembled on the genome of strain 287/91 (22). The *de novo* and reference contigs were combined using MUMmer (12) and Gap4 (5).

Sequences were annotated using GenoPipe (http://genopipe.bioinfo-portal.cdac.in/) and a combination of gene prediction software (1, 8, 18, 21). Manual curation followed to enhance the annotation, including pseudogene prediction and assignment of start sites. Genes with unsuitable names for submission were searched against SwissProt (23), and genes with a large degree of overlap were checked for domains (2, 9) and hits in SwissProt. If no domains or matches were found, the gene was removed from the annotation.

Intraserovar comparisons indicated that the complete *S*. Typhimurium 4/74 genome contained just eight single-nucleotide polymorphisms (SNPs) relative to SL1344, consistent with the shared history of the strains and high-quality sequencing and assembly. The *hisG* allele varied between the strains as expected (10).

Nucleotide accession numbers. Sequences were deposited in GenBank and assigned the following accession numbers; *S*. Typhimurium 4/74 (CP002487-CP002490), *S*. Choleraesuis SCSA50 (CM001062 to CM001063), *S*. Dublin SD3246 (CM001151 to CM001152), and *S*. Gallinarum SG9 (CM001153 to CM001154).

* Corresponding author. Mailing address: The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, United Kingdom. Phone: 44 131 651 9100. Fax: 44 131 440 0434. E-mail: Mark.Stevens@roslin.ed.ac.uk.

▽ Published ahead of print on 8 April 2011.

We gratefully acknowledge the support of the European Commission (EADGENE network of excellence, contract number FOODCT-2004-506416), the Biotechnology & Biological Research Council (core

strategic grants to The Roslin Institute and The Institute for Animal Health and India Partnering Award IPA1825) and the Department of Information Technology, Govt. of India.

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Bateman, A., et al.** 2004. The Pfam protein families database. Nucleic Acids Res. **32:**D138–D141.
3. **Bolton, A. J., G. D. Martin, M. P. Osborne, T. S. Wallis, and J. Stephen.** 1999. Invasiveness of Salmonella serotypes Typhimurium, Choleraesuis and Dublin for rabbit terminal ileum in vitro. J. Med. Microbiol. **48:**801–810.
4. **Bolton, A. J., M. P. Osborne, T. S. Wallis, and J. Stephen.** 1999. Interaction of Salmonella choleraesuis, Salmonella dublin and Salmonella typhimurium with porcine and bovine terminal ileum in vivo. Microbiology **145 (Pt 9):** 2431–2441.
5. **Bonfield, J. K., K. Smith, and R. Staden.** 1995. A new DNA sequence assembly program. Nucleic Acids Res. **23:**4992–4999.
6. **Chadfield, M. S., D. J. Brown, S. Aabo, J. P. Christensen, and J. E. Olsen.** 2003. Comparison of intestinal invasion and macrophage response of Salmonella Gallinarum and other host-adapted Salmonella enterica serovars in the avian host. Vet. Microbiol. **92:**49–64.
7. **Chiu, C. H., et al.** 2005. The genome sequence of Salmonella enterica serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. Nucleic Acids Res. **33:**1690–1698.
8. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
9. **Finn, R. D., et al.** 2010. The Pfam protein families database. Nucleic Acids Res. **38:**D211–D222.
10. **Hoiseth, S. K., and B. A. Stocker.** 1981. Aromatic-dependent Salmonella typhimurium are non-virulent and effective as live vaccines. Nature **291:**238–239.
11. **Jones, P. W., P. Collins, and M. M. Aitken.** 1988. Passive protection of calves against experimental infection with Salmonella typhimurium. Vet. Rec. **123:** 536–541.
12. **Kurtz, S., et al.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5:**R12.
13. **Li, R., et al.** 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. **20:**265–272.
14. **Paulin, S. M., A. Jagannathan, J. Campbell, T. S. Wallis, and M. P. Stevens.** 2007. Net replication of Salmonella enterica serovars Typhimurium and Choleraesuis in porcine intestinal mucosa and nodes is associated with their differential virulence. Infect. Immun. **75:**3950–3960.
15. **Paulin, S. M., et al.** 2002. Analysis of Salmonella enterica serotype-host specificity in calves: avirulence of S. enterica serotype gallinarum correlates with bacterial dissemination from mesenteric lymph nodes and persistence in vivo. Infect. Immun. **70:**6788–6797.
16. **Pullinger, G. D., et al.** 2007. Systemic translocation of Salmonella enterica serovar Dublin in cattle occurs predominantly via efferent lymphatics in a cell-free niche and requires type III secretion system 1 (T3SS-1) but not T3SS-2. Infect. Immun. **75:**5191–5199.
17. **Rankin, J. D., and R. J. Taylor.** 1966. The estimation of doses of Salmonella typhimurium suitable for the experimental production of disease in calves. Vet. Rec. **78:**706–707.
18. **Schattner, P., A. N. Brooks, and T. M. Lowe.** 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. **33:**W686–W689.
19. **Smith, H. W.** 1955. Observations on experimental fowl typhoid. J. Comp. Pathol. **65:**37–54.
20. **Stevens, M. P., T. J. Humphrey, and D. J. Maskell.** 2009. Molecular insights into farm animal and zoonotic Salmonella infections. Philos Trans. R. Soc. Lond. B Biol. Sci. **364:**2709–2723.
21. **Suzek, B. E., M. D. Ermolaeva, M. Schreiber, and S. L. Salzberg.** 2001. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics **17:**1123–1130.
22. **Thomson, N. R., et al.** 2008. Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res. **18:**1624–1637.
23. **Uniprot Consortium.** 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. **39:**D214–D219.
24. **Wallis, T. S., S. M. Paulin, J. S. Plested, P. R. Watson, and P. W. Jones.** 1995. The Salmonella dublin virulence plasmid mediates systemic but not enteric phases of salmonellosis in cattle. Infect. Immun. **63:**2755–2761.
25. **Watson, P. R., S. M. Paulin, A. P. Bland, P. W. Jones, and T. S. Wallis.** 1995. Characterization of intestinal invasion by Salmonella typhimurium and Salmonella dublin and effect of a mutation in the invH gene. Infect. Immun. **63:**2743–2754.
26. **Watson, P. R., S. M. Paulin, P. W. Jones, and T. S. Wallis.** 2000. Interaction of Salmonella serotypes with porcine macrophages in vitro does not correlate with virulence. Microbiology **146:**1639–1649.
27. **Wray, C., and W. J. Sojka.** 1978. Experimental Salmonella typhimurium infection in calves. Res. Vet. Sci. **25:**139–143.
28. **Yu, H., et al.** 2006. Complete nucleotide sequence of pSCV50, the virulence plasmid of Salmonella enterica serovar Choleraesuis SC-B67. Plasmid **55:** 145–151.

BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 1. 1–12
Advance Access published on 9 March 2012 doi:10.1093/bib/bbs007

The automatic annotation of bacterial genomes

Emily J. Richardson and Mick Watson

Submitted: 30th September 2011; Received (in revised form): 4th February 2012

Abstract

With the development of ultra-high-throughput technologies, the cost of sequencing bacterial genomes has been vastly reduced. As more genomes are sequenced, less time can be spent manually annotating those genomes, resulting in an increased reliance on automatic annotation pipelines. However, automatic pipelines can produce inaccurate genome annotation and their results often require manual curation. Here, we discuss the automatic and manual annotation of bacterial genomes, identify common problems introduced by the current genome annotation process and suggests potential solutions.

Keywords: *bacteria; genomics; annotation; automatic; errors*

BACKGROUND

Prokaryotic genomics has seen an explosion in the number of genome projects, driven by the advent of next generation sequencing (NGS), resulting in a huge reduction in the time and money investment per project [1]. Microbial genome annotation often consists of running an automatic annotation pipeline followed by manual curation of the results [2]. Most annotation pipelines use homology methods to transfer information from a closely related reference genome to the new sequence. Automatic pipelines can lead to the introduction and propagation of poor annotation and errors, and it is the purpose of the manual curation step to catch and remove these. However, as it is now possible to sequence multiple microbial genomes in a single day at low cost using a single sequencing machine [3], it is no longer feasible to manually curate the annotation of all sequenced genomes. Fully-automatic annotation pipelines, while essential to the modern microbial genomicist, may introduce and propagate inconsistent and incorrect gene annotations.

High-quality annotation goes beyond applying gene prediction software and transferring the annotation from the genome's closest relative. We have to include features other than coding sites (CDS), such as ribosomal-binding sites (RBSs), termination sites and conserved motifs/domains. Not only do these features give a fuller annotation they actually can rectify errors from earlier parts of the annotation process. For example, predicting RBS and termination sites will give a much clearer idea of a gene's true location rather than using gene prediction alone. Luckily, there are many software tools for the prediction of these features [4–8].

Transferring annotation purely based on the closest annotated relative does have its limitations. When we consider the reason the new strain has been sequenced, often it will be to identify how this strains differ genetically to its close relatives. This is paradoxical because we are trying to find the differences between these strains but using a similarity based method to annotate it. Potential areas of interest may not be annotated because they are not in the reference genome.

With this surge in sequencing, we will also see an increase in the number of annotated genomes submitted to the public databases. Sequence databases have introduced more stringent requirements for submitters meaning that running an annotation

Corresponding author. Mick Watson, ARK-Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, EH25 9RG, UK. Tel: +44 (0)131 651 9100; Fax: +44 (0)131 651 9105; E-mail: mick.watson@roslin.ed.ac.uk

Emily Richardson is a PhD student at The Roslin Institute, University of Edinburgh. Her project focuses on the use of next-generation bioinformatic and informatic tools for the multi-dimensional annotation of bacterial genomes.

Mick Watson is Director of ARK-Genomics, a genomics facility at the The Roslin Institute, University of Edinburgh. His research interests are the bioinformatics and functional genomics of farmed animal species and their pathogens.

pipeline alone is not enough to ensure acceptance of the genome annotation [9, 10]. There has also been a surge in other next generation techniques such as RNA-seq, incorporating experimental methods gives a better indication of a protein's role and whether it is functional. These annotations would be more accurate because they are based on actual experiment data rather than homology. Currently genomes can include evidence tags stating how the annotation was assigned, however, they are often omitted from the process. Including evidence qualifiers gives the user an idea of the reliability of the reference genome. The concept of assigning a level of quality to annotation is not novel, but is seldom used [11, 12].

This article discusses some of the current steps for prokaryotic genome annotation and offers a guide to some of the common problems that are encountered during automatic annotation. It goes on to identify the limitations of reference genomes and why choosing the closest relative is not always the best option. We also discuss the rules of the public sequence databases, and go on to suggest possible next steps toward a more accurate, comprehensive annotation with minimal propagation of errors.

Annotation of bacterial genomes

Here we describe a very general process used for bacterial genome annotation (Figure 1). A more thorough review can be found in Stothard and Wishart [2]. In many cases there is a closely related strain/serovar available which has already been sequenced and annotated. Most annotation pipelines employ gene prediction software, the most common of which is GLIMMER [13]. This uses a reference set of sequences to train a model and then utilizes that model to predict coding regions in the genome of interest. Many other *ab initio* gene prediction algorithms exist and these are reviewed by Do and Choi [14]. Alternatively, gene finding can be performed by extrinsic methods, identifying open reading frames directly from comparisons to protein databases [15, 16].

Once coding regions have been identified, they are aligned either to a reference genome annotation or the entirety of UniProt [17] using fast sequence alignment tools (e.g. FASTA [18] or BLAST [19]), the top hits are accepted as homologs and the annotation is transferred across for genes displaying high similarity. Other features such as tRNAs and rRNAs may then added using other prediction software [20].

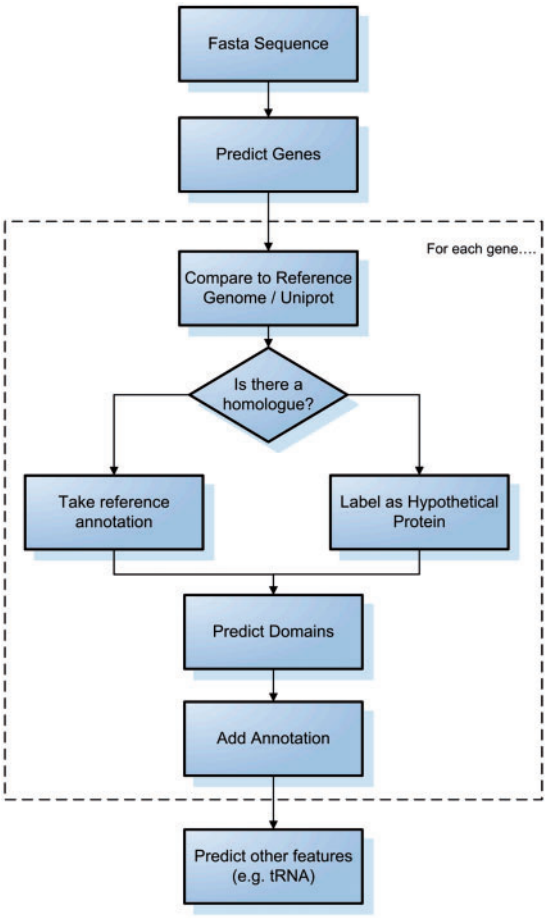


Figure 1: A generic process for bacterial genome annotation.

A range of automatic bacterial annotation pipelines have been published, including web-based systems such as RAST [21], BASys [22], WeGAS [23] and MaGe/Microscope [24]; and systems to be locally installed, such as AGeS [25], DIYA [26] and PIPA [27]. There is also MICheck [28] which checks annotated sequences for syntactic errors. All of these systems carry out the basic process outlined above, with various additions to check for errors or add additional information. It is worth noting that in order to submit to a genome repository that the annotation needs to be in a compatible format (e.g. .tab or .asn). Some pipelines do not output in this manner as they are designed to either hold the annotation online or for in-house analysis [22, 23]. Further processing may therefore be necessary before submission to a public database.

Other feature types

For acceptance to databases such as GenBank or EMBL, only gene, CDS and structural RNA features

need to be added [9, 10]. However, many other features should be added. This section gives a broad overview of some of the other features and how they can be predicted; a comprehensive guide is available [29].

Gene prediction software sometimes assigns the wrong start/termination sites. Glimmer for example assigns the start site as the most upstream start codon [5]. By searching for RBS, one can infer and reassign the start site; RBSFinder does this by looking for motifs such as the Shine-Dalgarno sequence pattern [5]. For termination sites, TransTerm searches for rho-independent transcription terminators to assign the correct termination site [6]. As well as correcting start/termination sites these features should be added to the annotation, using the tags 'RBS' and 'terminator' respectively.

Regions of conservation within proteins such as motifs and domains should be added to the annotation after the gene finding step. There are many databases which store protein families such as ProSite, PRINTS and Pfam [4, 7, 8]. InterproScan can perform searches against a range of domain/motif databases [30]. Hits to motif/domain databases should be assigned the qualifier 'db_xref' within the corresponding CDS feature [9, 10].

Areas of horizontal gene transfer (HGT) such as pathogenicity islands and prophage can be predicted by looking at asymmetries in codon composition and the GC content as these will often differ between areas of HGT and the rest of the genome [31]. They are often associated with the presence of integrases, transposases and IS elements [31]. Software tools exists to predict these [32, 33], and these are reviewed and compared by Langille, *et al.* [34]. There are clear guidelines for annotating phage, this should be assigned under the 'source' feature with the name of the bacteriophage in the 'organism' qualifier and the type of sequence in 'mol_type' (usually genomic DNA). There is no specific annotation tag for other GIs so these should be annotated as miscellaneous features. The mobile genetic elements themselves use the 'mobile_element' tag.

Sequence repeats such as 'clustered regularly interspaced short palindromic repeats' (CRISPRs) and other tandem repeats are of biological interest. For example, they can be used to understand the bacterial defense mechanism [35] and to distinguish between closely related strains [36]. Software tools exist [37, 38] and databases such as MICdb store predicted microsatellites as well as offering a prediction tool for user inputted sequence [39].

Identifying a protein's cellular localization can be indicative of function and this can be used in the identification of drug targets. There are many methods of prediction including homology and keywords [40], amino acid composition [41–43] and a mixture of these [44], Gardy and Brinkman [45] have performed a comprehensive review of the many tools available.

LIMITATIONS OF THE ANNOTATION PROCESS

In an ideal world this would be the end of the annotation process. The fact that homology is the basis for these pipelines means that many genomes currently available may have been annotated using old, out of date genomes as a reference which in turn have been annotated based on even older more out of date genomes. The misannotations and errors may perpetuate throughout each new genome, ultimately propagating into secondary databases such as UniProt [17] and KEGG [46], and domain-specific databases such as PATRIC [47].

The public sequence databases have recognized the need for controlling this replication of errors and provide validation software for checking the standard of one's annotation prior to submission [9, 10]. This section looks at common errors that are the product of automated annotation and tries to address methods of overcoming these.

Inconsistent annotation

Many bacterial genera now have multiple species and strains with complete genomes, representing a fantastic resource for comparative genomics. However, each genome is annotated separately, by a range of different groups using different protocols, and this introduces inconsistencies. One particular problem is that of split/fused genes and domains; Kummerfield and Teichman [48] found that, of 7116 distinct domain architectures examined across 131 archaeal, bacterial and eukaryotic genomes, 47% showed evidence of gene fusion/fission events. An example of this is the eutM/eutN locus in *Salmonella*. Figure 2 shows six different models that have been used to annotate this region in the 17 RefSeq records for *Salmonella* at time of publication. In *Salmonella typhi* CT18 (NC_003198) and *Salmonella typhi* Ty2 (NC_004631) there is a single ORF of 690 bp

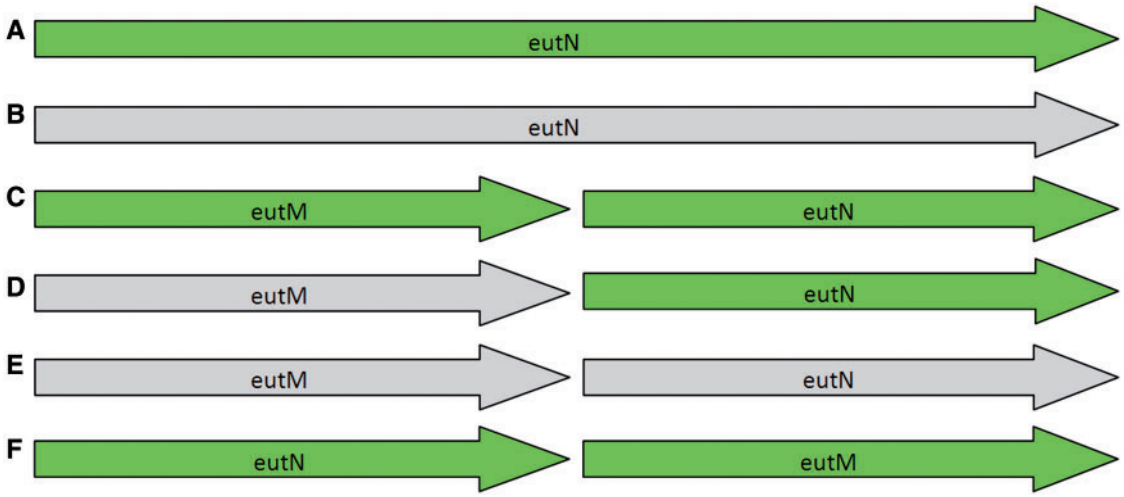


Figure 2: The six different models present across 17 RefSeq entries for Salmonella species for the eutM/eutN locus. Green indicates normal gene/CDS features, lighter grey indicates gene features annotated as pseudogenes. (A) A single intact gene of 690 bp; (B) a single pseudogene of 690 bp; (C) two short intact genes ~300 bp in length; (D) one pseudogene and one intact gene, each ~300 bp in length; (E) two pseudogenes, each 300 bp in length; and (F) two intact genes with the order reversed.

annotated as eutN (Figure 2A). The protein sequence maps to two domains in PFAM, a BMC domain (PF00936) and a EutN_CcmL domain (PF03319). In all other *Salmonella* genomes in RefSeq, stop codons within this region split the gene, and the domains, in two. In one genome (NC_012125) the region has been annotated as a single long pseudogene of 690 bp (Figure 2B); a further four genomes annotate two intact gene/CDS features, eutM and eutN, each ~300 bp in length (Figure 2C). A further three genomes are annotated with one pseudogene, a 291 bp ORF equivalent to the eutM gene in Figure 2C, and one intact gene, a 288 bp ORF labeled as eutN (Figure 2D). A further two genomes annotate two ORFs, 291 bp and 300 bp in length respectively, both annotated as pseudogenes (Figure 2E), equivalent to the eutM and eutN genes in Figure2C. Finally, one genome (NC_006511) includes two intact genes, but has reversed the order of eutM and eutN (Figure 2F).

The various ways in which the eutN and eutM genes have been annotated represents a problem for further genome annotation. We cannot know, simply from the genome sequences alone, whether this locus represents a single long gene that has been split in two, or two shorter genes that have become fused. All six models represent different interpretations of a locus that is highly conserved at the nucleotide level across *Salmonella* species, and any novel genome that is compared to just one of those models will have annotation heavily influenced by that model. For example, if a novel genome is compared only to genomes represented by Figure 2B (two short ORFs annotated as a single long pseudogene) the interpretation will be very different than if the genome were compared to Figure 2C (two short ORFs annotated as two separate intact genes).

Predicting domains directly, rather than genes, using tools such as PfamAlyzer [49], may help in regions with split genes. In the case of eutM/eutN in *Salmonella*, a domain search would identify two intact domains in all cases; however, the question of whether or not those domains come from the same or separate genes would remain unresolved. We are left with two different versions of the eutN gene from *Salmonella* in the public databases, one of 690 bp containing two domains, and one of ~290 bp with one domain.

The only way to annotate this region correctly *in silico* would be to compare any new genome to each of the six different models. It is difficult to imagine a set of rules that could be given to an automatic annotation pipeline to interpret correctly the evolution of this region and apply that interpretation to a newly sequenced genome. To truly get the full story we would need to look at experimental data (such as RNA-Seq data) to see what the patterns of expression are.

In the eutN/eutM example above, we see a case where genes of vastly differing lengths have been

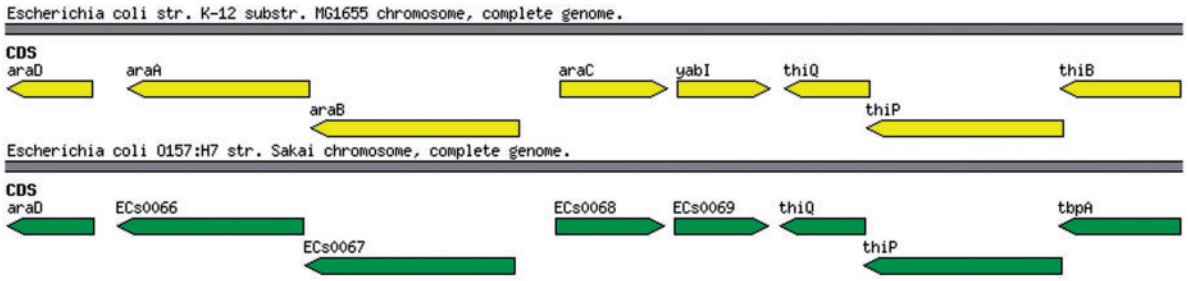


Figure 3: A syntenic block of genes showing inconsistent gene name annotations in *E.coli KI2 MGI655* and *E. coli 0I57:H7 Sakai*.

given the same gene name in different genomes; in contrast to this, it is also possible for orthologous genes to be assigned different gene names. Figure 3 shows a syntenic block of genes annotated in *Escherichia coli K12 MG1655* (NC_000913) and *E. coli O157:H7 Sakai* (NC_002695). These two regions are more than 97% identical at the nucleotide level; however, the annotation differs considerably. While *E. coli K12 MG1655* contains features with gene names araA, araB and araC, the equivalent features in *E. coli 0157:H7 Sakai* do not have those gene names and have been assigned uninformative locus tags. Further information is available for the features with only locus tags, including their involvement in arabinose metabolism, however, the gene names remain absent. At the far right of the gene block, two orthologous features exist, both with gene names, however, this time the problem is that they are different: thiB in *K12 MG1655* and tbpA in *0157:H7 Sakai*. A simple search of the NCBI gene database (search term 'thiB AND Escherichia coli [Organism]' versus search term 'tbpA AND Escherichia coli [Organism]') reveals that both features code for a thiamin(e) transporter subunit, but the gene is given the gene name tbpA in over 30 *E. coli* species, whereas it is given the name thiB in only one. Luckily, the thiB feature in *K12 MG1655* lists tbpA as a 'synonym'. Finally, in the centre of the image, *K12 MG1655* contains a feature with the gene name yabI, whereas its ortholog in *0157:H7 Sakai* only has a locus tag. This is an example of a y-gene, which we discuss in greater detail in the 'Hypothetical proteins' section.

The major issue here is that not only do different genomes annotate orthologous genes differently, and provide inconsistent information; they also contain differing amounts of information. This means that, when annotating a new genome, it is essential to choose a reference genome that contains the most accurate and up-to-date information, and that it is also preferable to compare any new genome to multiple references such that inconsistent annotations can be identified and resolved.

Spelling mistakes

There are 128 proteins in UniProt that contain the word 'syntase', an incorrect spelling of the word 'synthase'. To put this into context, the RefSeq entry for *Rhizobium etli CFN 42* (accession NC_007761) assigns the function 'dihydrofolate syntase' to gene folC. This has propagated into other databases such as UniProt (accession: Q2KE79), KEGG (accession: RHE_CH00024), and xBASE (accession: RHE_CH00024). If a user was to visit any of these databases and search for 'dihydrofolate synthase' the misspelled entries would be omitted from the search results. Large scale detection and correction of spelling mistakes in public databases is a difficult task, and so there is a reliance on the submitter to correct these. Automatic annotation pipelines simply copy and propagate what is there already. Spelling mistakes may be highlighted by the validation software provided by the public databases during submission, however, an alternative correct spelling isn't offered, making it difficult to amend the mistakes without manual intervention.

This can be solved by writing rules to find spelling mistakes [16]. However, this approach is limited to spelling mistakes which are explicitly written in the code. A solution may exist beyond biological science. The search engine Google upon receiving the input 'syntase' automatically states 'Did you mean: *synthase*'. There are programming languages which have classes or plugins to produce such 'did you mean' results [50, 51].

'Same gene name, different product name'

This issue occurs when two features, either within or between genomes, are assigned the same short gene name yet different product names. The NCBI validation software specifically highlights when this occurs intra-genomically with the description 'Same gene name, different product name' [9, 10]. In the current set of 2696 microbial genome and plasmid sequences in RefSeq, we detected 23,843 genes with at least two different product names (see http://www.ark-genomics.org/genomeannotation.html for the full list). The most extreme example of this is gene 'tnp' which has 151 different product names ('tnpA' has a further 97). A more manageable example can be seen in Table 1. The 'int' gene has a total of 12 different product names across 17 *Salmonella* RefSeq entries. These product names contain huge variation in terms of information content. When using an automatic annotation pipeline, there is a danger that if the top hit is to an entry labeled 'Hypothetical protein', then you will capture far less information than if your top hit is to 'phage integrase family site specific recombinase'. In order to correctly annotate this gene in a new genome, it is necessary to take into account all of these product names in the annotation process. It is difficult to imagine a set of text-mining rules that could efficiently interpret the range of annotations and assign the most suitable one to a new gene.

Hypothetical proteins

The term 'hypothetical protein' often refers to a gene that has been predicted by software but which finds no homolog of known function in the databases, and which has no known functional domain. There are currently 53 035 proteins whose product name contains both words in UniProt (search term: 'name:hypothetical AND name:protein') and there are a further 5 178 212 proteins in UniProt that contain the words 'uncharacterized' and 'protein' (search term: 'name:uncharacterized AND name:protein'). These may be real genes with no known function or they may be artifacts of the gene prediction process.

Many bacterial genes of unknown function are assigned y-gene names based on their orthologous location in *E. coli K-12* [52]. The letters denote the location in terms of minutes around a circular genome. This gene annotation has propagated throughout many strains and species of bacteria, losing the relevance and context of its name as the genes are not all in the same relative location to the original annotation in *E. coli K-12*. For example the *yabF* gene has a known function, 'glutathione-regulated potassium-eflux system ancillary protein'. The gene name *yabF* is completely meaningless in all genomes other than the original and actually has a synonym *kefF*. With that in mind annotators should use more informative gene names as a preference, choosing alternative gene names over the original y-gene annotation.

Often there are features which are only orthologous to other hypothetical features and do not contain any domains. These could either be regions with no functionality, a relic of the feature prediction software or the domains present have not been discovered yet. Whether or not to include them is often a decision made by the annotation team and varies between groups. Thus, many artifactual

Table 1: Different product names assigned to features with the gene name '*int*' across 17 different RefSeq entries for *Salmonella* species

Gene name	Product name	Accession
int	bacteriophage integrase	NC.003198, NC.004631, NC.015761
int	Gifsy-1 prophage Int	NC.006905
int	hypothetical protein	NC.006905
int	Integrase	NC.003198, NC.004631, NC.006511, NC.012125
int	integrase (fragment)	NC.003198
int	phage integrase family site specific recombinase	NC.006905
int	putative cytoplasmic protein	NC.006905
Int	Putative integrase	NC.003384
int	putative integrase protein	NC.006905
int	putative P4-type integrase	NC.006905
int	putative phage integrase protein	NC.006905
int	site-specific recombinase, phage integrase family	NC.012125

'hypothetical proteins' may be annotated, published and disseminated into the public databases, reinforcing the annotator's belief that their new gene predictions do indeed have homologs in other species. It would be more informative to actually state in the annotation a score for each feature. This will allow users to make informed assessments of the features and programmers to easily parse genomes to handle hypothetical proteins based on their quality of hits. Gilks, *et al.* [12] discuss the possibility of assigning scores based on the source of annotation.

There are arguments for and against keeping these proteins in the annotation. If they are indeed a misannotation by the gene prediction software they should be removed as they will perpetuate through secondary and tertiary databases as a recognized protein awaiting functional discovery. Searching for conserved domains or motifs in databases such as Pfam or InterPro can give an indication of whether a hypothetical protein is functional but this has pitfalls too. The fact that a protein has a domain hit doesn't necessarily convey its function. Pfam [8], for example, contains over 3000 'domains of unknown function', or DUFs, representing over 20% of known families [53] and as more novel genomes are sequenced the number of new DUFs will increase. A hit to a DUF does not inform us of a feature's function, but as they are areas of high conservation they indicate a potential region of biological interest.

Through computational methods alone there are no means to conclusively determine whether a genomic region is functional. With that in mind conserved features of unknown function should be kept because in the future they may be recognized as a true region of interest; however, they should be annotated differently to discriminate them from features with stronger evidence. Evidence tags are available but they are often not present, and are not a prerequisite for submission to GenBank or Embl. Evidence qualifiers such as how the feature was predicted (e.g. glimmer, blast, homology) and what entries it hits in a given database provide a clear audit trail for anyone who wants to assess the quality of a particular annotation. The type of data source used, that is, whether it is manually curated or automated should be stated, providing the user with a clear method of judging the annotation. As experimental data becomes more ubiquitous evidence tags should play a larger role in annotation.

Experimental methods such as RNA-Seq [54] and Signature Tagged Mutagenesis (STM) [55] may help to identify regions of functionality. RNA-Seq data can help delineate and quantify areas of transcription, and overlaying this expression data on the genome may help biologists to identify pseudogenes and the true locations of features. STM can help identify the function of genes by monitoring the phenotype of single-gene mutants.

The most important point is that one's annotation is only ever as good as the reference data sources. In terms of publicly available genome sequences the quality is varied. It is worth actually looking at the annotation and assessing the quality. Choosing a genome because it is the closest relative will give the most homologous features but might not give the best quality annotation.

Combining additional data with the original annotation gives scientists a new way of viewing the genome. Experimental data could be able to solve the *eutM/eutN* problem described above; for example, RNA-Seq data would show which areas of the genome are actively transcribed and STM may indicate whether knocking out either of the genes alters the phenotype of the mutant.

Distinguishing orthologs from paralogs

The definition of orthologous and paralogous genes is of great importance when annotating novel genomes. Whereas 'homology' refers to genes that simply share a common origin, 'orthology' refers to genes that arise by speciation and 'paralogy' refers to genes that arise by duplication. Figure 4 shows some of the processes that can lead to, and define, orthologs and paralogs. Beginning with a single ancestral, a gene duplication event occurs to create two paralogous genes. After a speciation event, there are two different organisms that both contain the paralogous genes from the gene duplication event. Gene 1a in Organism 1 has three homologs after the speciation event. Gene 1a in Organism 1 and Gene 1a in Organism 2 are orthologs as they have only been separated by the speciation event. Gene 1a in Organism 1 and Gene 1b in Organism 1 are in-paralogs, as they have only been separated by the gene duplication event. Finally, Gene 1a in Organism 1 and Gene 1b in Organism 2 are out-paralogs, as they have been separated by the gene duplication and the speciation event.

These processes are not only crucial in defining evolutionary relationships, but also functional

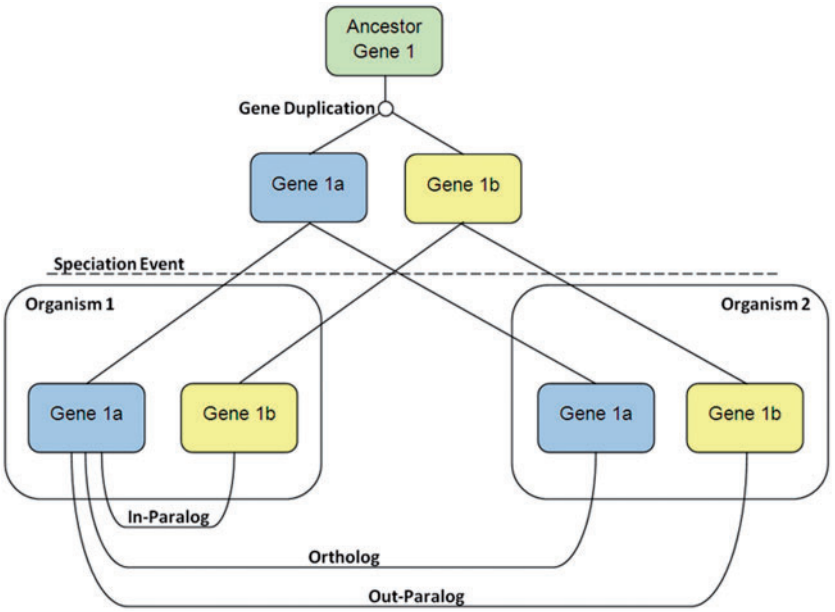


Figure 4: A diagram displaying the processes that can lead to, and define, orthologs and paralogs. Gene duplication and speciation events create complex evolutionary relationships between genes.

relationships, as orthologs tend to retain similar functions, whereas paralogs tend to diverge over time to perform different functions (reviewed in ref. [56]). Therefore, when transferring functional annotation from a sequenced genome to a novel genome, it is essential that orthologs are accurately defined. There are several computational approaches which can be used to accurately define orthologs (reviewed in ref. [57]). Phylogenetic tree-based approaches attempt to reconstruct the evolutionary relationship between gene sequences and thus define orthologs and paralogs; however, it may be impractical to construct a phylogenetic tree for every gene in a newly sequence genome. An alternative is the 'bidirectional' or 'reciprocal' best-hit approach [58], usually determined by comparing the top-ranking matches found by a search algorithm such as BLAST or FASTA [18, 19]. Gene Synteny, the conservation of local gene order, can also help distinguish orthologs from paralogs in closely related genomes. However, it is important to note that a number of processes can lead to the breakdown of absolute gene synteny, resulting in genuine orthologs having a different gene order. These processes include gene duplication or fusion events, local rearrangements (insertions/deletions) and translocations. It is important that we model these processes to allow the correct identification of orthologs in complex cases, and the MaGe [24] system attempts to do this. Finally, it has been observed that orthologs exhibit a greater level of protein domain architecture conservation than paralogs [59]. In practice, it may be essential to use a combination of approaches, and several software applications exist [57].

THE RULES OF THE SEQUENCING DATABASE

Many scientists go through the process of annotation with the final aim of submitting to a genome database such as GenBank or EMBL. In order to realize this goal there are many rules which need to be followed [9, 10] and often validation software is provided to verify one's annotation. These rules are imposed to ensure a better standard of genome annotation, however, they do mean that often the output of an automatic annotation pipeline must be manually checked and altered prior to publication. Many of the issues described in the 'Limitations of the Annotation Process' section may be identified as potential problems and the submitter is provided with long lists of features that represent these. They must be checked, and either altered or justified. In addition to those mentioned above, there are others described below.

CDS nomenclature

There are many words which may be unacceptable in protein names, such as 'binding', 'domain', 'like', 'motif', 'gene' and 'homolog'. Submitters may be

encouraged to change these: for example 'bacteriophage replication gene' can be changed to 'bacteriophage replication protein' and 'peptidyl-tRNA hydrolase domain protein' can be changed to 'peptidyl-tRNA hydrolase protein'; a note may be added to state that the feature contains the aforementioned domain. These rules add complications if the submitter wants to fully automate the process of annotation. As a rule of thumb, if a predicted coding region has homologs in SwissProt these are the best protein names to transfer across and running the validation software after using SwissProt initially can greatly reduce the number of suspect names. As an aside, 'probable' and 'predicted' are not flagged up by the validation software but 'putative' is the preferred alternative.

Some CDSs have the same protein name as the protein next to them, which can be the sign of either a disrupted gene or a valid gene duplication event. It can also be because the protein name is very general such as 'hypothetical protein' or 'inner membrane protein'. These features may be flagged up by the validation software and, if they are not pseudogenes, need a note stating that they overlap a CDS with the same protein name.

CDS gene names that appear more than once in a genome and have different proteins names to one another (e.g. Table 1) may also be identified as potential errors. These may be brought to the submitter's attention who often has to use their discretion and knowledge to assign gene names correctly. This can be as simple as performing a similarity search and seeing which gene names are associated with the hits.

Problems with coding regions

The NCBI validation software flags up all instances where a coding region completely contains another coding region on the opposite strand. The submitter is asked to check these coding regions and decide whether these are true features. If the coding region only hits hypothetical proteins and doesn't contain any domains, it may be either removed or demoted to a miscellaneous feature.

FUTURE

Gold standard genomes

RefSeq is one attempt to standardize and improve the quality of genome annotation; however, as we have shown, problems persist. With the implementation of stricter rules for submission we should see an increase in annotation quality. While genomes of varying quality are available there should be a means for scientists to see the quality of any given annotation. Evidence qualifiers such as how the feature was predicted and what entries in a given database the feature sequence hit, including the database version and date, would provide a clear audit trail for anyone who wants to assess the quality of a particular annotation. The type of data source used, that is, whether it is manually curated or automatically generated should also be stated, providing the user with a clear method of judging the annotation.

Out of the 1851 publicly available completed bacterial genomes 102 have a version number of 0.2 or higher [60]. This means that the submitting group have revisited the original sequence and changed it. The fact that the sequences have been changed is indicative of a higher quality sequence. This, however, does not reflect the quality of the annotation. It is possible to look at the revision history of genomes within GenBank, this will give users an idea of changes on a genome by genome basis, no small feat when there are 1851 genomes available. In the literature there have been several papers which have revisited and reannotated genomes, these include strains of *E. coli, Campylobacter jejuni* and *Mycobacterium tuberculosis* [61–63]. In terms of what is currently available these genomes are likely to be the closest to realizing 'gold standard genome annotation'.

Janssen, *et al.* [11] calculated the number of publications per gene for all completed genome to calculate a Species Knowledge Index (SKI) for each genome. They showed that, in bacteria, there is a pronounced bias toward certain organisms namely *E. coli*, *Pseudomona aeruginosa* and *Bacillus subtilis*. With this in mind perhaps there should be a focus to annotate genomes with a high SKI to the highest level possible as there is such an abundance of experimental data available. These can then be used as gold standard genomes for annotations of other species.

As we learn more about genes and protein function it becomes clear that a simple protein name is inadequate. Some proteins are multi-functional, performing different tasks depending on the context it is expressed in. We can say that a protein has a one-to-many relationship with function, meaning that assigning a protein name based on the first function associated with it can be misleading and

inaccurate. The Gene Ontology (GO) may provide a more flexible way of describing a range of functions explicitly and concisely, and GO annotations natively include evidence qualifiers. However, GO terms are not frequently included as part of the initial annotation of bacterial genomes. The EBI offer UniProtKB-GOA Proteome Sets [64], GO annotations for all completely sequenced genomes in the public domain, however, these are not included with or clearly linked to the original genome submission. The development and use of GO annotations is encouraged and these should be included in genome annotation efforts.

Improving automated annotation

The pipelines currently on offer do not take many of the pitfalls outlined above into account, meaning that a lot of manual effort is required to correct errors and inconsistencies. It is easy to imagine adjustments to current pipelines that take into account certain aspects (e.g. common spelling mistakes) but not others (e.g. correctly interpreting pseudogenes). Realistically, completely removing the manual stage of annotation would be imprudent, however, improving current automated pipelines may greatly reduce the time spent manually checking the annotation.

New data types

There have been a flood of new genome-wide data types in the post-genomic era, for example microarray and RNA-Seq data, many of which can assist with genome annotation. However, these are often large, unwieldy, come in a variety of different formats and can be hard to integrate with one another. Allowing scientists to visualize this data alongside genome annotation can be hugely powerful [65]; however, genome annotation is often kept in specific flat file formats where integrating non-text data is virtually impossible. Secondary and tertiary databases may include additional data alongside the original genome annotation [20], but these 'data warehouse' approaches employ copies of the original data which can become out-of-date and out-of-synch with the original data. The advent of bioinformatics web services [66] may allow new systems that query data live over the internet, ensuring the latest data is displayed.

CONCLUSION

Advances in sequencing technologies are allowing researchers to sequence microbial genomes at a huge rate. It is becoming harder to devote time to manually annotate these genomes, leading to a rise in automatic annotation pipelines. However, due to a range of problems, the output of these automatic annotation pipelines is unsuitable for publication. Some changes can be made to improve this output; however, it is difficult to envisage an end to manual checking and curation.

Additional data from post-genomics experiments can help improve genome annotation; however, a line has to be drawn regarding what data should be included in the annotation and what should be in separate databases. Tools and services need to be developed which offer scientists a means of viewing genome annotation augmented with other experimental data. This will empower the user to make meaningful judgments on the quality of annotation and the relevance of a particular region to their research.

For the foreseeable future bacterial annotation requires both automated and manual steps. Offering users a measure of quality for the whole genome and individual genes will allow user to make an informed choice regarding reference genomes and transferring annotation between genomes. Using GO terms would improve protein description and reduce syntactic errors.

Key Points

- Advances in sequencing technology now allow modern researchers to rapidly sequence multiple bacterial genomes.
- Automatic annotation pipelines that work via comparison to a reference database can introduce and propagate errors.
- Manual checking and curation of annotation is essential to maintain a high quality.
- Additional data-sources from post-genomic experiments can assist in the annotation process.

Acknowledgements

We would like to acknowledge the help, assistance and advice provided by staff at the NCBI and EBI during genome submission.

FUNDING

This work was funded by an Institute Strategic Programme grant awarded to The Roslin Institute by the Biotechnology and Biological Sciences

Research Council (BBSRC), and by a studentship funded by the Institute for Animal Health.

References

1. MacLean D, Jones JD, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol* 2009;**7**:287–96.
2. Stothard P, Wishart DS. Automated bacterial genome analysis and annotation. *Curr Opin Microbiol* 2006;**9**:505–10.
3. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;**11**:31–46.
4. Attwood TK, Bradley P, Flower DR, *et al*. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 2003; **31**:400–2.
5. Suzek BE, Ermolaeva MD, Schreiber M, *et al*. A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics* 2001;**17**:1123–30.
6. Ermolaeva MD, Khalak HG, White O, *et al*. Prediction of transcription terminators in bacterial genomes. *J Mol Biol* 2000;**301**:27–33.
7. Sigrist CJ, Cerutti L, de Castro E, *et al*. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 2010;**38**:D161–166.
8. Finn RD, Mistry J, Tate J, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2010;**38**:D211–22.
9. The Bacterial Genome Submission Guide. http://www.ncbi.nlm.nih.gov/genbank/genomesubmit.html (25 November 2011, date last accessed).
10. Genome Project Submission Account guidelines. http://www.ebi.ac.uk/embl/Submission/genomes.html (25 November 2011, date last accessed).
11. Janssen P, Goldovsky L, Kunin V, *et al*. Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. *EMBO Rep* 2005;**6**: 397–9.
12. Gilks WR, Audit B, de Angelis D, *et al*. Percolation of annotation errors through hierarchically structured protein sequence databases. *Math Biosci* 2005;**193**:223–34.
13. Delcher AL, Harmon D, Kasif S, *et al*. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999;**27**:4636–41.
14. Do JH, Choi DK. Computational approaches to gene prediction. *J Microbiol* 2006;**44**:137–44.
15. Frishman D, Mironov A, Mewes HW, *et al*. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* 1998;**26**:2941–7.
16. Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* 1999; **16**:512–24.
17. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* 2011;**39**:D214–9.
18. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* 1990;**183**:63–98.
19. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
20. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997;**25**:955–64.
21. Aziz RK, Bartels D, Best AA, *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008;**9**:75.
22. Van Domselaar GH, Stothard P, Shrivastava S, *et al*. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res* 2005;**33**:W455–9.
23. Lee D, Seo H, Park C, *et al*. WeGAS: a web-based microbial genome annotation system. *Biosci Biotechnol Biochem* 2009;**73**: 213–6.
24. Vallenet D, Labarre L, Rouy Z, *et al*. MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* 2006;**34**:53–65.
25. Kumar K, Desai V, Cheng L, *et al*. AGeS: a software system for microbial genome sequence annotation. *PLoS One* 2011;**6**:e17469.
26. Stewart AC, Osborne B, Read TD. DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinform* 2009; **25**:962–3.
27. Yu C, Zavaljevski N, Desai V, *et al*. The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. *BMC Bioinform* 2008;**9**:52.
28. Cruveiller S, Le Saux J, Vallenet D, *et al*. MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. *Nucleic Acids Res* 2005;**33**:W471–9.
29. Webin EMBL-EBI annotation features and qualifiers. http://www.ebi.ac.uk/ena/WebFeat/ (25 November 2011, date last accessed).
30. Mulder N, Apweiler R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* 2007;**396**:59–70.
31. Hacker J, Blum-Oehler G, Mühldorfer I, *et al*. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol* 1997;**23**: 1089–97.
32. Hsiao W, Wan I, Jones SJ, *et al*. IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinform* 2003;**19**: 418–20.
33. Waack S, Keller O, Asper R, *et al*. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinform* 2006;**7**:142.
34. Langille M, Hsiao W, Brinkman F. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinform* 2008;**9**:329.
35. Barrangou R, Fremaux C, Deveau H, *et al*. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 2007;**315**:1709–12.
36. Kassai-Jáger E, Ortutay C, Tóth G, *et al*. Distribution and evolution of short tandem repeats in closely related bacterial genomes. *Gene* 2008;**410**:18–25.
37. Bland C, Ramsey TL, Sabree F, *et al*. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinform* 2007;**8**:209.
38. Grissa I, Vergnaud G, Pourcel C, *et al*. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 2007;**35**:W52–7.
39. Sreenu VB, Alevoor V, Nagaraju J, *et al*. MICdb: database of prokaryotic microsatellites. *Nucleic Acids Res* 2003;**31**: 106–8.

40. Lu Z, Szafron D, Greiner R, *et al*. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* 2004;**20**:547–56.
41. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. *Bioinform* 2001;**17**:721–8.
42. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci* 2004;**13**:1402–6.
43. Wang J, Sung W-K, Krishnan A, *et al*. Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC Bioinform* 2005;**6**:174.
44. Gardy JL, Laird MR, Chen F, *et al*. PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* 2005;**21**:617–23.
45. Gardy JL, Brinkman FSL. Methods for predicting bacterial protein subcellular localization. *Nat Rev Micro* 2006;**4**: 741–51.
46. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
47. Snyder EE, Kampanya N, Lu J, *et al*. PATRIC: the VBI PathoSystems Resource Integration Center. *Nucleic Acids Res* 2007;**35**:D401–6.
48. Kummerfeld SK, Teichmann SA. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet* 2005;**21**:25–30.
49. Hollich V, Sonnhammer EL. PfamAlyzer: domain-centric homology search. *Bioinformatics* 2007;**23**:3382–3.
50. Lucene – java based search engine. http://lucene.apache.org/java/docs/index.html (25 November 2011, date last accessed).
51. PHP class – 'did you mean?'. http://www.phpclasses.org/package/4569-PHP-Get-spelling-correction-suggestions-from-Google.html (25 November 2011, date last accessed).
52. Rudd KE. Linkage map of Escherichia coli K-12, edition 10: the physical map. *Microbiol Mol Biol Rev* 1998;**62**: 985–1019.
53. Bateman A, Coggill P, Finn RD. DUFs: families in search of function. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2010;**66**:1148–52.
54. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
55. Saenz HL, Dehio C. Signature-tagged mutagenesis: technical advances in a negative selection method for virulence gene identification. *Curr Opin Microbiol* 2005;**8**:612–9.
56. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 2005;**39**:309–38.
57. Kristensen DM, Wolf YI, Mushegian AR, *et al*. Computational methods for Gene Orthology inference. *Brief Bioinform* 2011;**12**:379–91.
58. Overbeek R, Fonstein M, D'Souza M, *et al*. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 1999;**96**:2896–901.
59. Forslund K, Pekkari I, Sonnhammer EL. Domain architecture conservation in orthologs. *BMC Bioinform* 2011;**12**:326.
60. NCBI Complete Microbial Genomes. www.ncbi.nlm.nih.gov/genomes/lproks.cgi (25 November 2011, date last accessed).
61. Luo C, Hu GQ, Zhu H. Genome reannotation of Escherichia coli CFT073 with new insights into virulence. *BMC Genomics* 2009;**10**:552.
62. Gundogdu O, Bentley SD, Holden MT, *et al*. Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. *BMC Genomics* 2007;**8**:162.
63. Camus JC, Pryor MJ, Medigue C, *et al*. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology* 2002;**148**:2967–73.
64. Barrell D, Dimmer E, Huntley RP, *et al*. The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* 2009;**37**:D396–403.
65. Watson M. ProGenExpress: visualization of quantitative data on prokaryotic genomes. *BMC Bioinform* 2005;**6**:98.
66. Bhagat J, Tanoh F, Nzuobontane E, *et al*. BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res* 2010;**38**:W689–94.

BMC Microbiology

Research article

Analysis of the role of 13 major fimbrial subunits in colonisation of the chicken intestines by *Salmonella enterica* serovar Enteritidis reveals a role for a novel locus

Debra J Clayton*[1], Alison J Bowen†[1], Scott D Hulme†[1,3], Anthony M Buckley[1], Victoria L Deacon[1], Nicholas R Thomson[2], Paul A Barrow[1,3], Eirwen Morgan[1], Michael A Jones[1,3], Michael Watson[1] and Mark P Stevens[1]

Address: [1]Division of Microbiology, Institute for Animal Health, Compton, Berkshire, RG20 7NN, UK, [2]Pathogen Genomics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK and [3]School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington, Leicestershire, LE12 5RD, UK

Email: Debra J Clayton* - debra.clayton@bbsrc.ac.uk; Alison J Bowen - alisonbowen9@hotmail.com; Scott D Hulme - scott.hulme@nottingham.ac.uk; Anthony M Buckley - tony.buckley@bbsrc.ac.uk; Victoria L Deacon - victoria.deacon@bbsrc.ac.uk; Nicholas R Thomson - nrt@sanger.ac.uk; Paul A Barrow - paul.barrow@nottingham.ac.uk; Eirwen Morgan - eirwen.morgan@bbsrc.ac.uk; Michael A Jones - michael.a.jones@bbsrc.ac.uk; Michael Watson - michael.watson@bbsrc.ac.uk; Mark P Stevens - mark-p.stevens@bbsrc.ac.uk

* Corresponding author †Equal contributors

Published: 18 December 2008

BMC Microbiology 2008, **8**:228 doi:10.1186/1471-2180-8-228

Received: 15 September 2008
Accepted: 18 December 2008

This article is available from: http://www.biomedcentral.com/1471-2180/8/228

Abstract

Background: *Salmonella enterica* is a facultative intracellular pathogen of worldwide importance. Over 2,500 serovars exist and infections in humans and animals may produce a spectrum of symptoms from enteritis to typhoid depending on serovar- and host-specific factors. *S.* Enteritidis is the most prevalent non-typhoidal serovar isolated from humans with acute diarrhoeal illness in many countries. Human infections are frequently associated with direct or indirect contact with contaminated poultry meat or eggs owing to the ability of the organism to persist in the avian intestinal and reproductive tract. The molecular mechanisms underlying colonisation of poultry by *S.* Enteritidis are ill-defined. Targeted and genome-wide mutagenesis of *S.* Typhimurium has revealed conserved and host-specific roles for selected fimbriae in intestinal colonisation of different hosts. Here we report the first systematic analysis of each chromosomally-encoded major fimbrial subunit of *S.* Enteritidis in intestinal colonisation of chickens.

Results: The repertoire, organisation and sequence of the fimbrial operons within members of *S. enterica* were compared. No single fimbrial locus could be correlated with the differential virulence and host range of serovars by comparison of available genome sequences. Fimbrial operons were highly conserved among serovars in respect of gene number, order and sequence, with the exception of *safA*. Thirteen predicted major fimbrial subunit genes were separately inactivated by lambda Red recombinase-mediated linear recombination followed by P22/int transduction. The magnitude and duration of intestinal colonisation by mutant and parent strains was measured after oral inoculation of out-bred chickens. Whilst the majority of *S.* Enteritidis major fimbrial subunit genes played no significant role in colonisation of the avian intestines, mutations affecting *pegA* in

two different *S*. Enteritidis strains produced statistically significant attenuation. Plasmid-mediated *trans*-complementation partially restored the colonisation phenotype.

Conclusion: We describe the fimbrial gene repertoire of the predominant non-typhoidal *S. enterica* serovar affecting humans and the role played by each predicted major fimbrial subunit in intestinal colonisation of the primary reservoir. Our data support a role for PegA in the colonisation of poultry by *S*. Enteritidis and aid the design of improved vaccines.

Background

Non-typhoidal serovars of *Salmonella enterica* are an important cause of food-borne diarrhoeal illness in humans worldwide. Using active surveillance data from a catchment area of 44.5 million people, the FoodNet network has estimated that there are 1.4 million cases of human non-typhoid salmonellosis in the United States per annum, leading to 15,000 hospitalisations and 400 deaths [1]. Over the past three decades *S. enterica* serovar Enteritidis has emerged as a significant cause of such infections [2]. The consumption of undercooked poultry meat and eggs is a major risk factor for *S*. Enteritidis infection [3] and the phage types circulating in humans are commonly found in broilers [4] and layers [5]. The incidence of *S*. Enteritidis infection in humans declined markedly following the implementation of control strategies, including vaccination for poultry, regulations on storage and preparation of food and improved education [6]. Despite such measures, *S*. Enteritidis remains the most prevalent cause of non-typhoidal salmonellosis in many countries, including the United Kingdom http://www.hpa.org.uk/infections/topics_az/salmonella/data.htm, and improved vaccines are needed to achieve further reductions in the burden of human disease.

It is well established that *S*. Enteritidis is able to persist in the intestinal and reproductive tract of poultry in the absence of clinical signs [7]; however the molecular mechanisms mediating colonisation of these sites are ill-defined. Further, it is unclear why some *S. enterica* serovars are associated with enteric disease in a broad range of healthy out-bred adult hosts (e.g. Enteritidis and Typhimurium), whereas others are host-restricted or -specific and associated with severe systemic disease (e.g. Gallinarum in poultry and Typhi in humans). Targeted and genome-wide mutagenesis of the broad host range serovar Typhimurium has indicated that it uses both conserved and host-specific factors to colonise the intestines of chickens, cattle, pigs and mice [8-14]. Among the factors that influence intestinal colonisation are fimbriae; proteinaceous surface appendages that mediate interactions between bacteria and host cells.

Of the thirteen fimbrial loci predicted to be encoded by the *S*. Typhimurium genome, *lpf*, *fim*, *bcf*, *stb*, *stc*, *std*, *sth* and *csg* have been implicated in virulence in mice [11,13,15-17]. Screening of a library of signature-tagged mutants of *S*. Typhimurium indicated that pathogenicity island (SPI)-6-encoded *saf* fimbriae may play a host-specific role in ileal colonisation of pigs [14], whereas the *stbC, csgD* and *sthB* fimbrial genes were implicated in colonisation of the avian gut [12]. Separately Ledeboer et al described a role for *lpfA-E*, *pefC*, *csgA* and *fimH*, but not *sthD* or *bcfF*, in biofilm formation on chicken intestinal mucosa cultured *ex vivo* [18]. Relatively few studies have probed the role of fimbriae in colonisation of poultry by *S*. Enteritidis. Allen-Vercoe and Woodward reported that a *S*. Enteritidis mutant lacking *fimD*, *csgA*, *pefC*, *lpfC* and *sefA* colonised the caeca at comparable levels to the parent strain following oral dosing of 1 or 5 day-old chicks [19] and was similarly invasive [20] and adherent to chicken gut explants [21]. Furthermore, single mutants lacking *fimA*, *csgA* or *sefA* exhibited no significant defect in colonisation of chick caeca and were excreted in the faeces at comparable levels to the parent [22,23]. Although roles for *S*. Enteritidis fimbriae in intestinal colonisation of poultry have so far been lacking, Type I fimbriae [24] and curli [25] have been implicated in egg contamination.

In the recent publication of the complete genome sequence of *S*. Enteritidis strain P125019 [26] we have defined the full repertoire of fimbrial loci and identified a unique fimbrial operon, *peg*, present in *S*. Gallinarum, *S*. Enteritidis and also *S*. Paratyphi. The *peg* operon displays 60–70% sequence conservation with the *stc* operon of *S*. Typhimurium and is located in the same relative position. The *peg* operon belongs to the γ clade of fimbriae and is predicted to be assembled via the chaperone usher pathway [27].

The work herein examined the fimbrial gene conservation in the published genomes of other *S. enterica* serovars and also searched for traits associated with phase variation. Isogenic *S*. Enteritidis mutants with insertions in the major fimbrial subunit of each of the fimbrial operons were constructed using lambda Red recombinase-mediated linear recombination [28] followed by P22/int transduction. Mutant phenotypes were then evaluated and confirmed using an established chicken colonisation model.

Methods

In silico *analysis of fimbrial loci*

The complete genome sequences of *S. enterica* serovar Enteritidis strain P125109 [26], *S.* Gallinarum strain 287/91 [26], *S.* Typhimurium SL1344 and *S.* Typhimurium DT104 were produced by the Pathogen Sequencing Unit, Wellcome Trust Sanger Institute, UK http://www.sanger.ac.uk/Projects/Salmonella/. Published genome sequences were obtained from the National Center for Biotechnology Information (NCBI) and are described with their RefSeq-curated accession numbers; *S.* Typhimurium LT2 NC_003197 [29], *S.* Typhi CT18 NC_003198 [30], *S.* Typhi Ty2 NC_004631 [31] and *S.* Choleraesuis SC-B67 NC_006905 [32]. Fimbrial gene sequences were identified from the primary literature and databases via NCBI Entrez and the genome sequences were visualised and compared using Artemis and Artemis Comparison Tool ACT [33,34]. Direct and indirect repeat sequences were searched for as described [35]. A Perl script was written to isolate and visualise *S.* Enteritidis fimbrial operons and is available from the authors on request.

Bacterial strains and plasmids

S. Enteritidis phage type 4 strain P125109 (NCTC 13349) was isolated from a poultry-associated outbreak in the UK and is naturally nalidixic acid resistant. A spontaneous nalidixic acid resistant derivative of *S.* Enteritidis S1400 [19] was selected by standard methods and it exhibits wild-type growth and chick colonisation phenotypes (data not shown). Strains were cultured in Luria-Bertani (LB) medium supplemented with antibiotics at the following concentrations where appropriate: nalidixic acid (Nal, 20 μg ml^{-1}), novobiocin (1 μg ml^{-1}), ampicillin (100 μg ml^{-1}) and chloramphenicol (25 μg ml^{-1}). Plasmids pCP20 [36], pKD3 and pKD46 [28] were obtained from the *E. coli* Genetic Stock Centre, Yale University. Plasmids pCR4Blunt-TOPO (Invitrogen, Paisley, UK) and pACYC177 [37] were used for cloning in *E. coli* K-12 strain TOP10F' (Invitrogen, Paisley, UK).

Construction and validation of major fimbrial subunit mutations

Primers were designed to amplify the pKD3-encoded chloramphenicol resistance cassette, including 40 bp homology extensions from the 5' and 3' of each predicted major fimbrial subunit gene (Table 1). The extensions were designed such that the region between the start and stop codon of each major fimbrial subunit gene would be replaced by the chloramphenicol resistance cassette. PCR products were purified and electroporated into *S.* Enteritidis harbouring the helper plasmid pKD46, following induction of the Red recombinase with 10 mM L-arabinose at 30°C as previously described [28]. Recombinants were selected on LB-agar containing chloramphenicol and cured of pKD46 by culture at 37°C in the absence of ampicillin. Mutations were confirmed at the expected position in the genome by PCR with primers specific to the chloramphenicol resistance cassette and primers flanking each major fimbrial subunit gene (Table 2). Mutations were also confirmed by Southern blotting with *Hind*III-digested genomic DNA from wild-type and mutant strains using the *cat* gene as a probe. Attempts were made to transduce each mutation using bacteriophage P22/int into an archived strain to reduce the likelihood that phenotypes are the result of second site defects. For unknown reasons, three mutations could not be transduced, therefore the original recombinant was compared relative to the parent strain. Growth kinetics of all mutants were determined by diluting an overnight culture of *S.* Enteritidis wild-type or mutant strain 1:1000 in LB medium and measuring the absorbance at 600 nm every 30 minutes for 24 hours using a Bioscreen-C real-time spectrophotometer (Thermo®, Helsinki, Finland).

FLP recombinase-mediated excision of the chloramphenicol resistance cassette

To remove the chloramphenicol resistance cassette from the Δ*pegA*::*cat* mutant and create a predicted non-polar mutation, the temperature-sensitive plasmid pCP20 was introduced and expression of FLP recombinase induced by culture at 42°C in the absence of antibiotic selection as described [28]. FLP-mediated recombination between flippase recognition target (FRT) sites flanking the pKD3-derived chloramphenicol resistance cassette was confirmed by PCR using primers flanking *pegA*. Excision of the chloramphenicol cassette was predicted to result in an 84 nucleotide in-frame scar between the *pegA* start and stop codons. The second codon in the scar is a stop codon, however *pegA* is not predicted to be translationally coupled to the 3' gene and therefore polar effects are not anticipated at the level of transcription or translation.

Trans-*complementation of the* pegA::cat *mutant*

The *pegA* coding sequence was amplified by PCR from *S.* Enteritidis P125109 genomic DNA using *Pfu* proof-reading DNA polymerase (Promega, Madison, USA) with primers *pegA*-for and *pegA*-rev containing *Cla*I restriction endonuclease cleavage sites. The *pegA* amplicon was ligated into pCR4Blunt-TOPO via topoisomerase I and transformed into chemically-competent *E. coli* TOP10 F' cells as described by the manufacturer (Invitrogen, Paisley, UK). A recombinant was verified by PCR with *pegA*-specific primers and digestion with *Cla*I. The *Cla*I fragment containing *pegA* was then sub-cloned into pACYC177 using T4 DNA ligase. Recombinant plasmids with the insert in the sense (p$pegA_{fwd}$) and antisense (p$pegA_{rev}$) orientation relative to the kanamycin resistance gene promoter of pACYC177 were isolated and electropo-

Table 1: Primers used to construct major fimbrial subunit mutations in *S.* Enteritidis P125109

Name	Sequence (5'-3')
stbAFmut	ATGTCTATGAAAAAATATTTAGCAATGATCACAGGCTCGCTGTGTAGGCTGGAGCTGCTTCG
stbARmut	TTATTTATACGAAACGGCGTATTGTAGGGTGGCAGCGACTCATATGAATATCCTCCTTA
pegAFmut	ATGAAACGTTCACTTATTGCTGCTTCTGTATTGTCTGCTGTGTGTAGGCTGGAGCTGCTTCG
pegARmut	TTAATCAGTTAATACCGTCATCGTCAGTACAGATTCAACACATATGAATATCCTCCTTA
stdAFmut	GTGCTTCGTTTAACACCAGGCGTTTATTATTCATACGAATTGTGTAGGCTGGAGCTGCTTCG
stdARmut	TCACAGGTATTTCAGGGTGTAGGTGACGGATGCGTTGAAGCATATGAATATCCTCCTTA
steAFmut	ATGAAGTCATCTCATTTTTGTAAACTGGCAGTAACTGCATGTGTAGGCTGGAGCTGCTTCG
steARmut	TTACAGGTAAGAGATAGTGACGTTGGCGGCGCTGCTGAACATATGAATATCCTCCTTA
stfAFmut	ATGAATACAGCAGTAAAAGCTGCGGTTGCTGCCGCACTGGTGTGTAGGCTGGAGCTGCTTCG
stfARmut	TTACAGATAGCTGATCGTGAAGTTTACGGTGCTGCTGAATCATATGAATATCCTCCTTA
sthAFmut	ATGTTTAATAAGAAAATTATCATCCTGGCAATGTTAACTTGTGTAGGCTGGAGCTGCTTCG
sthARmut	TTACTGATACGAAACGGTATACGTAACCTGAGTGCTAACACATATGAATATCCTCCTTA
stiAFmut	ATGAAACTCTCCTTAAAAACACTCACTGTGGCACTGCCGTGTGTAGGCTGGAGCTGCTTCG
stiARmut	TCAGTTATATTGCAGATAGAATGTTGCGGTTGCATCGACCCATATGAATATCCTCCTTA
bcfAFmut	ATGAAAAAGCCTGTACTAGCATTAATGGTCTCTGCCATTGTGTGTAGGCTGGAGCTGCTTC
bcfARmut	TCAGGAATAAACCATGCTAAATGTCGCCGTCGCGGTAACCATATGAATATCCTCCTTA
csgAFmut	ATGAAACTTTTAAAAGTGGCAGCATTCGCAGCAATCGTAGTTGTGTAGGCTGGAGCTGCTTCG
csgARmut	TTAATACTGGTTAGCCGTGGCGTTGTTGCCAAAACCAACCCATATGAATATCCTCCTTA
lpfAFmut	ATGGAGTTTTTAATGAAAAAGGTTGTTTTTGCTCTGTCTGTGTGTAGGCTGGAGCTGCTTCG
lpfARmut	TTATTCGTAGGACAGGTTGAAGTCACTTCTGCGTTACCGCATATGAATATCCTCCTTA
fimAFmut	ACCTCTACTATTGCGAGTCTGATGTTTGTCGCTGGCGCATGTGTAGGCTGGAGCTGCTTCG
fimARmut	TTATTCGTATTTCATGATAAAGGTGGCGTCGGCATTAGCCTGCATATGAATATCCTCCTTA
sefAFmut	ATGCGTAAATCAGCATCTGCAGTAGCAGTTCTTGCTTTAATGTGTAGGCTGGAGCTGCTTCG
sefARmut	GTTTTGATACTGCTGAACGTAGAAGGTCGCAGTGGGTCCATTTCATATGAATATCCTCCTTA
safAFmut	GTGGTTATTCAAATGAAAAGCATAAAAAAAATTGATTATCGTGTGTAGGCTGGAGCTGCTTCG
safARmut	TTAAGGCTGATATCCCACTACGTCTACAGTTATTGGGTACCATATGAATATCCTCCTTA

The primers were designed to mutate the major fimbrial subunit by lambda Red recombinase-mediated integration of linear PCR products. Forward and reverse primers were used to amplify the pKD3-derived chloramphenicol cassette and contain 40 bp homology extensions 5' and 3' of the target gene.

Table 2: Primer combinations used to validate each fimbrial mutation.

Primer combination	Predicted amplicon size (bp)	Sequence (5'-3')
bcfAFOR + C1	633	TGCACTATCCGCAACGATATATTT
bcfAREV + C2	507	TAAAATACGCTTTCGCGATCGGTCGGT
csgAFOR + C2	173	CAAGGAGCAATAAAGTATGCATAATTT
csgAREV + C1	302	CAGCAGTTGTAGTGCAGAAACAGTCGCATA
lpfAFOR + C2	867	TTAGTTACGCGCTGTGTCAA
lpfAREV + C1	288	ATCCAATACCCACCTCTATACACTCCA
fimAFOR +C1	807	AACCTCAGATCGCACCTGCTGC
fimAREV + C2	429	ATGCCGACATGACGCCAGACC
sefAFOR + C1	373	CTATTAATGGGGATGTTGTGTAA
sefAREV + C2	946	CTAATAATCTCTTATAATTTC
safAFOR + C1	701	TGAGACTCTCTCATTGGAGCGCT
safAREV + C2	597	AATTGAGGTCAAGGGTCGCGCC
stbAFOR + C2	887	TTAATGGTGGGGGACATCGTA
stbAREV + C1	295	TTATTTTTACCACTCCATAAGCACGAA
pegAFOR + C2	179	CACAAGCCAGGCATAATGCAATCATC
pegAREV + C1	377	ACATTGCGATAACTTCCTGTCTATGAGAA
stdAFOR + C2	587	GCTGTACCGTACCTGACTGTC
stdAREV + C1	714	TGTTTTTAAATTTCATCCGCGAAG
steAFOR + C1	739	TACGACAACGCCTATATAATA
steAREV + C2	600	AGCAGCGTGGAGTGTCCCAGGTCAGC
stfAFOR + C1	283	CATATAAACATGGGGTATTGATGA
stfAREV + C2	155	GGCTGGCATCATCTTTAACA
sthAFOR + C1	584	GCGTTGATTTTGTTAATGC
sthAREV + C2	704	GAAAGCTCACGATTTGAGATCAAC
stiAFOR + C2	385	TTTGGCCGACAACACACTATG
stiAREV + C1	661	GTAAATCAGCTTAAATTCCG
C1	-	TTATACGCAAGGCGACAAGG
C2	-	GATCTTCCGTCACAGGTAGG

Primers are specific to the flanking regions of the specified fimbrial gene (FOR or REV) or the chloramphenicol resistance cassette (C1 – forward or C2 – reverse).

rated into *S*. Enteritidis P125109 Δ*pegA*::*cat* with selection for ampicillin resistance.

Experimental animals

Inoculation of chickens with *S*. Enteritidis wild-type, mutant and *trans*-complemented strains was conducted according to the requirements of the Animal (Scientific Procedures) Act 1986 (PPL 30/1998) with the approval of the local Ethical Review Committee. Specific pathogen-free out-bred Rhode Island Red chickens were reared at the Institute for Animal Health and housed in group cages in bio-secure accommodation. Birds were fed a vegetable-based diet (Special Diet Services, Manea, Cambridgeshire, UK) with access to water *ad libitum*. To reduce inter-animal variation, chickens were orally dosed on the day of hatch with 0.1 ml *Salmonella*-free adult gut flora cultured as described [38]. Owing to constraints of space, the phenotype of each fimbrial mutant could not be simultaneously evaluated relative to the parent. Rather, 4 groups of 15 birds were accommodated per room with 3 groups each receiving a different fimbrial mutant strain and one group the corresponding parent strain. Approximately 1.5 × 10^8 colony-forming units (CFU) of stationary phase LB-grown *Salmonella* were given by oral gavage at 18-days-old. Inocula were confirmed to be comparable by retrospective plating of serial dilutions to selective media. Five birds from each group were sacrificed by cervical dislocation at 3, 7 and 10 days post-inoculation and the liver, spleen, caecal contents, caecal wall, ileal contents and ileal wall were recovered aseptically and diluted 1:10 in phosphate-buffered saline for homogenisation. A rotary blade was used to homogenise the samples and serial ten-fold dilutions were plated on brilliant green agar containing novobiocin and nalidixic acid. As each sample was diluted 1:10 for homogenisation and 20 μl of this was plated in triplicate, the theoretical limit of detection by direct plating is $\log_{10}$ 2.2 CFU/g. For some samples bacterial counts were below the limit of detection by direct plating and therefore enrichment was used. The homogenized sample was incubated overnight at 37°C in a final concentration of 1 × selenite broth before being plated on brilliant green agar plates supplemented with nalidixic acid and novobiocin. This results in a qualitative rather than a quantitative count but was given an arbitrary figure of $\log_{10}$ 1 CFU/g as the sample diluted 10^{-1} must have contained at least one viable organism. Owing to the difficulty separating caecal contents from the mucosa, the total caecal load is presented as a measure of colonisation of this site. This represents the mean viable count of *S*. Enteritidis in caecal content and mucosa samples, including biological and technical replicates.

To confirm the role of *pegA* in intestinal colonisation of chickens by *S*. Enteritidis, P125109 wild-type, Δ*pegA*::*cat* mutant, Δ*pegA* mutant, Δ*pegA*::*cat* [p$pegA_{fwd}$] and Δ*pegA*::*cat* [p$pegA_{rev}$] were given by oral gavage to ten 18-day-old Rhode Island Red chickens as above. Post mortem examinations were performed at 1 and 3 days post-inoculation (n = 5 per time interval) and bacteria at enteric and systemic sites enumerated. Plasmid stability in the absence of antibiotic selection *in vivo* was evaluated by plating selected samples to media containing nalidixic acid with or without ampicillin.

Statistical analysis

Counts of viable bacteria were $\log_{10}$ transformed and a generalised linear model was constructed using the least square means ± standard error of the mean (Statistical Analysis System version 9, SAS Institute, Cary, NC, USA). The significance of differences between test and control groups was determined by an F-test with data taken as repeated measurements. P values < 0.05 were considered significant.

Results

In silico *analysis of* S. *Enteritidis P125109 fimbrial loci*

Fourteen predicted fimbrial loci of the sequenced *S*. Enteritidis phage type 4 strain were identified [26]. Thirteen fimbrial loci are predicted to be encoded on the genome, whereas the P125109 *pef* operon is plasmid-encoded and highly similar to that of *S*. Typhimurium LT2 and *S*. Choleraesuis SC-B67. As *S*. Enteritidis *pef* was previously reported to play no significant role in colonisation of 1- and 5-day-old chicks [19], we elected to focus this study on chromosomally-encoded loci. Additional file 1 shows the predicted organisation of each fimbrial operon of strain P125109, together with %G+C content of the locus and the location and e-values of Pfam subunit, usher and chaperone domains.

The analysis of the Pfam domains failed to identify a major fimbrial subunit in *csg* and *saf*, consistent with the prediction that they give rise to atypical fimbriae. The *csg* operon is not predicted to encode proteins with usher or chaperone domains, consistent with assembly of Csg fimbriae via a nucleator-dependent pathway [39]. The *saf* operon consists of a chaperone and usher domain. The adhesive component is formed by the main structural subunit whose sequence has been shown here to be highly variable and it is not located at the tip as with other chaperone/usher assembled fimbriae [40]. The *saf* fimbriae are composed of flexible linear multi-subunit fibers connected by short fibers or linkers which allow flexibility in the final structure [40].

The conservation and organisation of fimbrial loci in the genomes of sequenced strains of *S. enterica* was analysed using the Artemis Comparison Tool. This revealed differences in the number and location of fimbrial loci between the strains as well as the presence of predicted truncations

and pseudogenes (Figure 1). At the nucleotide level, 56 of 71 fimbrial genes examined possessed ≥ 95% identity (Additional file 2). These include all of the genes of the fimbrial operons *sti*, *stb*, *fim*, *csg* and *lpf*, implying that their function may be conserved.

S. Enteritidis P125109 shares 10 of its 13 fimbriae with the sequenced *S*. Typhimurium strains. The *S*. Enteritidis *ste*, *sef* and *peg* operons are absent from the sequenced serovar Typhimurium strains, whereas the latter posses *stc* and *stj* operons that we do not find in P125109 (Figure 1).

We previously reported that the percentage of fimbrial genes that are pseudogenes in *S*. Gallinarum is greater than the genomic mean [26]. In addition we found here that the host-specific strains *S*. Typhi CT18 and *S*. Typhi Ty2 (data not shown) possessed the highest number of predicted fimbrial pseudogenes (based on the presence of at least one stop codon in the predicted coding sequence). The percentage of fimbrial genes that are pseudogenes in *S*. Typhi CT18 and *S*. Typhi Ty2 is 14% and 16% respectively, whereas the genomic mean of pseudogenes is 4.4%. In contrast the broad host-range serovars Enteritidis and Typhimurium LT2, DT104 and SL1344 appeared to contain an intact repertoire of fimbriae (data not shown) and the host-restricted serovar Choleraesuis maintained an intermediate number of predicted functional fimbrial genes. No single fimbrial locus could be correlated with host-specificity; however as has previously been suggested it is plausible that the loss of fimbrial genes in host-specific and -restricted serovars is associated with the narrowing of the niches they may occupy [26,29,30].

Fimbrial genes in some bacteria are subject to phase variable (on-off) expression that may be mediated via recombination (e.g. FimBE-mediated inversion of the *fimA* promoter in *E. coli* [41]), epigenetic regulation dependent on Dam methylation (e.g. control of Pap pili in uropathogenic *E. coli* [42]) or slipped-strand mis-pairing between homo- or hetero-polymeric tracts (e.g. assembly and maturation of Neisserial pilin reviewed in [43]). In *Salmonella*, evidence exists for phase variable expression of Type I fimbriae [44-46] and long-polar fimbriae [47]. Further, epigenetic regulation of the *pef* genes in *S*. Typhimurium by Dam methylation has been described [48] and expression of *std* fimbrial genes has been observed to be repressed in a *S*. Typhimurium Dam methylase mutant [49,50].

We searched *S*. Enteritidis fimbrial loci for traits associated with phase variation. Genes with homology to known recombinases were not detected within or proximal to fimbrial loci. Putative transposase and integrase genes associated with DNA mobility were observed proximal to the *saf*, *sef* and *fim* operons. Direct or inverted repeat sequences that may serve as substrates for recombination were not detected. A pattern-matching search was carried out for the Dam methylase target sequence GATC within and proximal to P125109 fimbrial operons. This identified hundreds of potential targets (Additional file 3), including those predicted to be methylated in the *S*. Typhimurium *pef* cluster [48]. Where *S*. Typhimurium strains SL1344 and LT2 possess GATC sites at -98, -110 and -212 relative to the start of *pefB*, *S*. Enteritidis P125109 possessed only the sites at -110 and -212, but an additional site at +47 in *pefB* that is absent in the two Typhimurium strains (Additional file 3a, grey shaded area). Three potential Dam methylation target sites were also identified upstream of the *std* operon (-88, -97 and -110) in *S*. Enteritidis P125109. This density of GATC sites is higher than random distribution would predict and correlates with the Dam-dependent repression of the *std* genes as detected by microarray analysis [49] and using antibody against StdA [50] (Additional file 3a, purple shaded area). Predicted Dam methylase targets were also identified upstream of the *sef*, *sti* and *stf* operons in *S*. Enteritidis P125109 (Additional file 3). Hundreds of homopolymeric tracts comprising 4 or more A or C residues were identified within fimbrial loci. Several conserved hetero-polymeric tracts were identified using a variable tandem repeat pattern finder, however only one was present in a fimbrial gene (ten repeated 6-mers (GACCAT) within *stdA*).

Construction of S. *Enteritidis major fimbrial subunit mutants*

Amplicons for the 13 chromosomally-encoded predicted major fimbrial subunit genes of *S*. Enteritidis P125109, were produced in order to delete each one via lambda Red recombinase-mediated linear recombination. Despite repeated attempts, pKD46 failed to mediate homologous recombination of linear amplicons in *S*. Enteritidis P125109 under conditions suitable for other *S. enterica* strains. However all 13 genes were successfully disrupted in the *S*. Enteritidis phage type 4 strain S1400nal^R, which is known to efficiently colonise the avian intestines [19,20]. Ten of the major fimbrial subunit gene deletions (marked by insertion of a chloramphenicol resistance cassette between the predicted start and stop codons) were successfully transduced into *S*. Enteritidis P125109 using bacteriophage P22/int. Transductants of *S*. Enteritidis P125109 or the archived S1400nal^R strain were not isolated for three of the mutated fimbrial constructs (Δ*safA*::*cat*, Δ*fimA*::*cat* and Δ*steA*::*cat)*. All of the successfully recovered isogenic mutants were verified by PCR and no growth defects were detected in batch culture (data not shown).

Screening of S. *Enteritidis fimbrial subunit mutants in a chick colonisation model*

Although P125109 is known to colonise the intestines of streptomycin pre-treated mice [51], no data existed on the colonisation dynamics of the sequenced *S*. Enteritidis

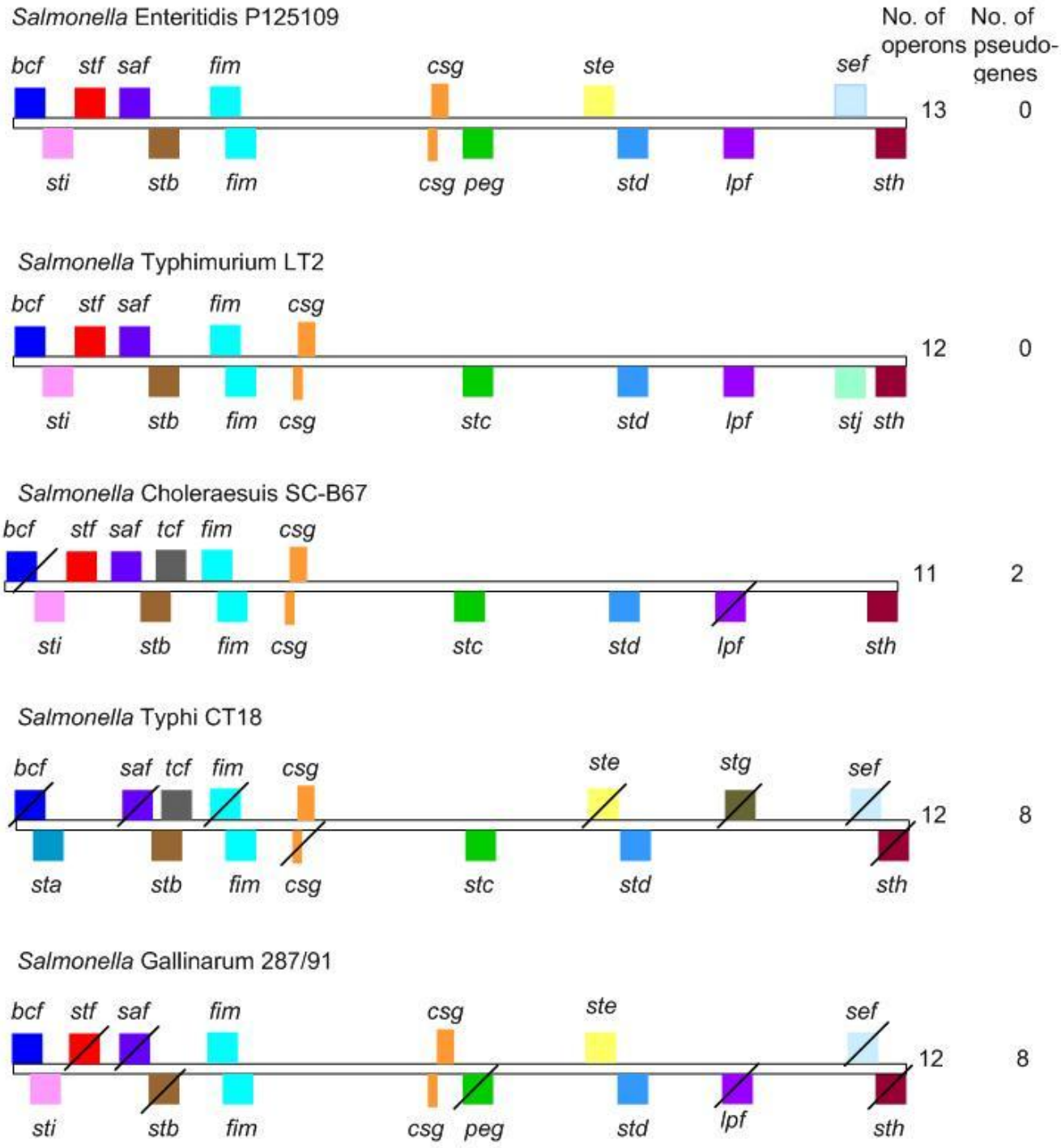


Figure 1

Schematic representation of the repertoire and relative genomic location of fimbrial loci in the published genomes of *S. enterica* serovars. Each coloured box represents a distinct fimbrial locus encoded in the sense (top) or anti-sense (bottom) orientation. Boxes of the same colour on both strands represent divergently transcribed operons. A diagonal line through the box indicates that at least one gene in the operon is a predicted pseudogene. The repertoire of *S.* Typhimurium and *S.* Typhi is representative of other sequenced strains of the same serovar. All genomes are aligned relative to their predicted origin. Not to scale.

strain P125109 in chickens. A pilot experiment was therefore performed to evaluate the magnitude and duration of colonisation of enteric and systemic sites at intervals post-oral inoculation and to gain an assessment of inter-animal variation. Following oral gavage of 18-day-old outbred specific pathogen-free Rhode Island Red chickens birds with 1.5×10^8 CFU, the caecal contents and mucosa were colonised by 4–5 $\log_{10}$ CFU/g of strain P125109 at days 1, 3 and 7 post-infection (n = 5 per time interval; Additional file 4). Bacterial colonisation of the ileum and translocation to the liver and spleen was detected, with recoveries at around the limit of detection at 2–3 $\log_{10}$ CFU/g by days 3 and 7 (Additional file 4).

Each fimbrial mutant was separately inoculated into groups of 15 Rhode Island Red chickens at 18 days-of-age and bacteria enumerated at enteric and systemic sites at 3, 7 and 10 days-post inoculation relative to the corresponding parent strain. As the caeca are a key site of bacterial persistence in the avian gut [52,53], (Additional file 4) and attenuation of defined and random *Salmonella* mutants is reliably detected at this site [12], the total caecal load is presented here as a measure of intestinal colonisation, representing the mean of the caecal wall and mucosa bacterial counts. We cannot preclude the possibility that some fimbriae mediate a specific tissue tropism that was not detected herein. Recoveries of viable bacteria from the liver and spleen were often close to the limit of detection by direct plating in chickens infected with wild-type strains (Additional file 4) and mutant strains (data not shown). Where adequate bacteria were recovered to permit a statistical analysis, no significant differences were observed at these sites. Figure 2 shows the caecal colonisation kinetics of *stb*,*peg*, *std*, *stf*, *sth*, *sti*, *bcf*, *csg*, *lpf* and *sef* major fimbrial subunit mutants of *S*. Enteritidis P125109 relative to the parent strain. The caecal loads of *fim*, *saf* and *ste* major fimbrial subunit mutants of *S*. Enteritidis strain S1400nal^R relative to the parent are shown in Figure 3.

The *S*. Enteritidis P125109 Δ*stbA*::*cat* fimbrial mutant was recovered from the chicken caeca at lower levels than the wild-type at all intervals post-inoculation (Figure 2A), with differences becoming significant by days 7 and 10 (P = 0.0081 and P = 0.03, respectively). This is consistent with the attenuation of a *S*. Typhimurium *stbC* mutant in chick caeca detected by signature-tagged mutagenesis [12]. The Δ*pegA*::*cat* mutant of *S*. Enteritidis P125109 was significantly impaired in colonisation of the caeca at days 3 and 7 post-inoculation compared with the wild-type (P = 0.0006 and P = 0.0002 respectively), although recoveries by day 10 were comparable (Figure 2B). The P125109 Δ*bcfA*::*cat* mutant, was recovered in significantly lower numbers than the parent strain at day 7 (P = 0.04), but not at other times (Figure 2G) and the S1400nal^R Δ*steA*::*cat* mutant was recovered in significantly lower numbers than the parent but only at day 10 (P = 0.0034; Figure 3C). No other fimbrial mutations significantly influenced the course of caecal colonisation at the 95% confidence interval.

Confirmation of the role of PegA in colonisation of chickens by S. Enteritidis

Figure 2B implies a role for PegA in the colonisation of the chicken caeca. However, the Δ*pegA*::*cat* mutation was transduced from S1400nal^R into P125109 prior to analysis in chickens and a theoretical possibility exists that other traits proximal to the *pegA* gene were transferred that resulted in attenuation. To address this, we analysed the phenotype of the original S1400nal^R Δ*pegA*::*cat* mutant relative to the parent and sought to restore the mutant to the wild-type level of colonisation by plasmid-mediated *trans*-complementation using the same experimental design as above. As with the Δ*pegA*::*cat* mutant of P125109 (Figure 2B) an approximate 2 $\log_{10}$ CFU/g reduction in the total caecal load of the S1400nal^R Δ*pegA*::*cat* mutant was detected at days 3 and 7 post-inoculation relative to the parent strain (Figure 4; P = 0.0135 and P = 0.0088, respectively). However, as with the Δ*pegA*::*cat* mutant of strain P125109, no significant difference was detected by day 10 post-inoculation.

The chloramphenicol resistance cassette was excised from the P125109Δ*pegA*::*cat* fimbrial mutant to determine if polar effects on the expression of 3' genes may explain the attenuation observed. This addresses the possibility that *pegA* may not be involved in colonisation *per se*, but that downstream genes participate in the expression of surface structure(s) that may include distally-encoded fimbrial subunits. Excision was achieved by transient expression of flippase recombinase as described in the Methods. The total caecal loads of both the *S*. Enteritidis P125109 Δ*pegA*::*cat* andΔ*pegA* mutant were approximately two orders of magnitude lower than the parent strain at 1 and 3 days post-oral inoculation of chickens (Figure 5). No significant difference existed between the caecal loads of the Δ*pegA*::*cat* and Δ*pegA* mutants (P values 0.27 and 0.64 at 1 and 3 days post-inoculation, respectively); however in both cases a highly significant reduction in caecal load was detected for each mutant relative to the parent strain (P values < 0.0001 at 1 day post-infection), consistent with previous findings.

A pACYC177-derived plasmid was created in which the *S*. Enteritidis P125109 *pegA* gene was cloned in the same orientation as the kanamycin promoter (p$pegA_{fwd}$), or in the antisense orientation (p$pegA_{rev}$). This replicon was selected for *trans*-complementation as it did not impair the virulence of *S*. Typhimurium in mouse co-infection studies [54]. Introduction of p$pegA_{rev}$ into the *S*. Enteri-

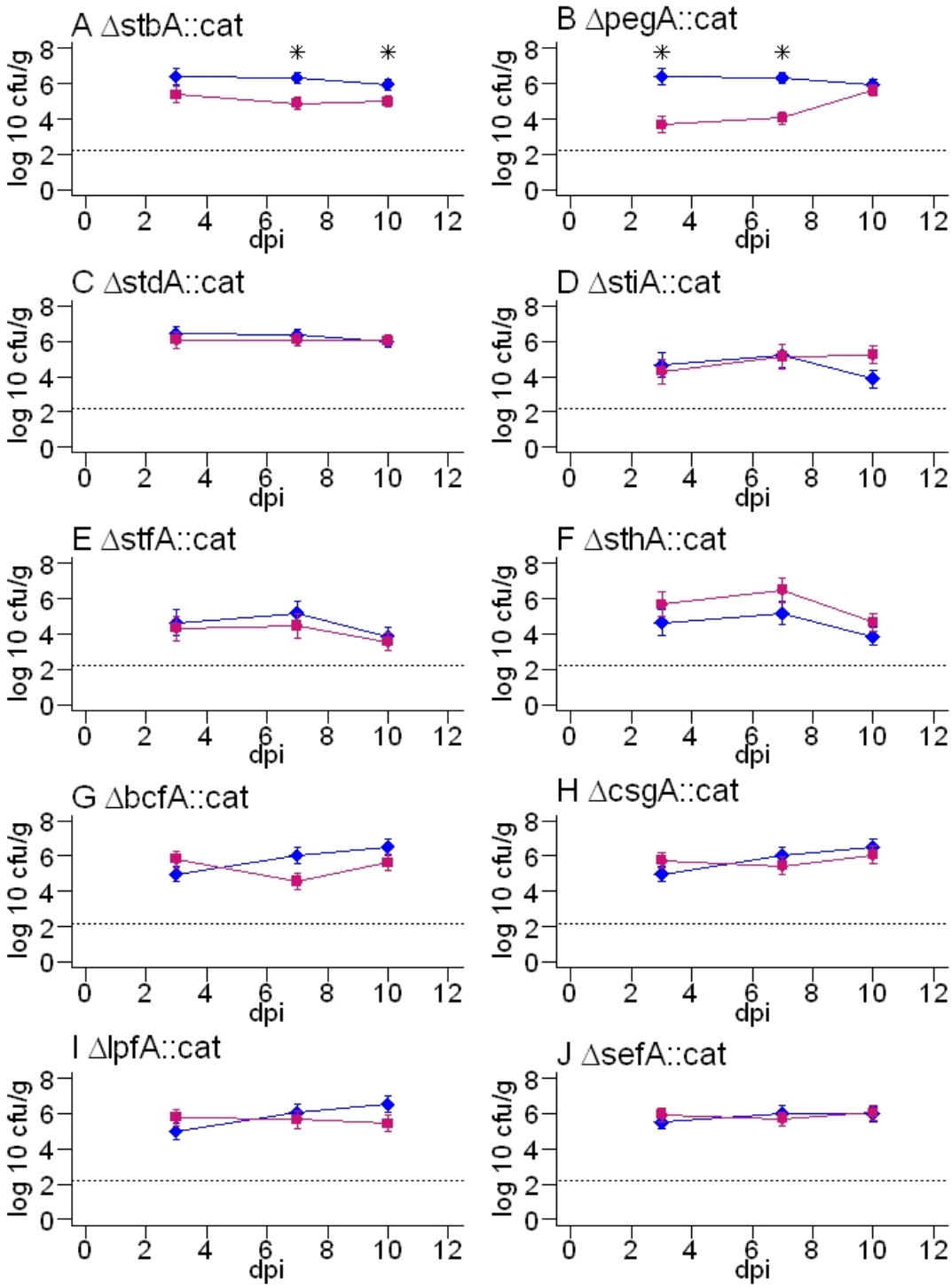


Figure 2

Total caecal load of *S*. Enteritidis P125109 wild-type and major fimbrial subunit mutant strains at 3, 7 and 10 days post-oral inoculation of 18-day-old out-bred Rhode Island Red chickens. Blue lines with diamonds denote the wild-type strain and pink lines with squares denote the fimbrial mutants. The dashed line indicates the theoretical limit of detection by direct plating (2.2 $\log_{10}$ CFU/g). The data reflect the mean ± standard error of the mean from five birds at each time interval. F-tests of the difference in recovery of wild-type and mutant strains at each time interval were performed and *P* values < 0.05 are marked with an asterisk.

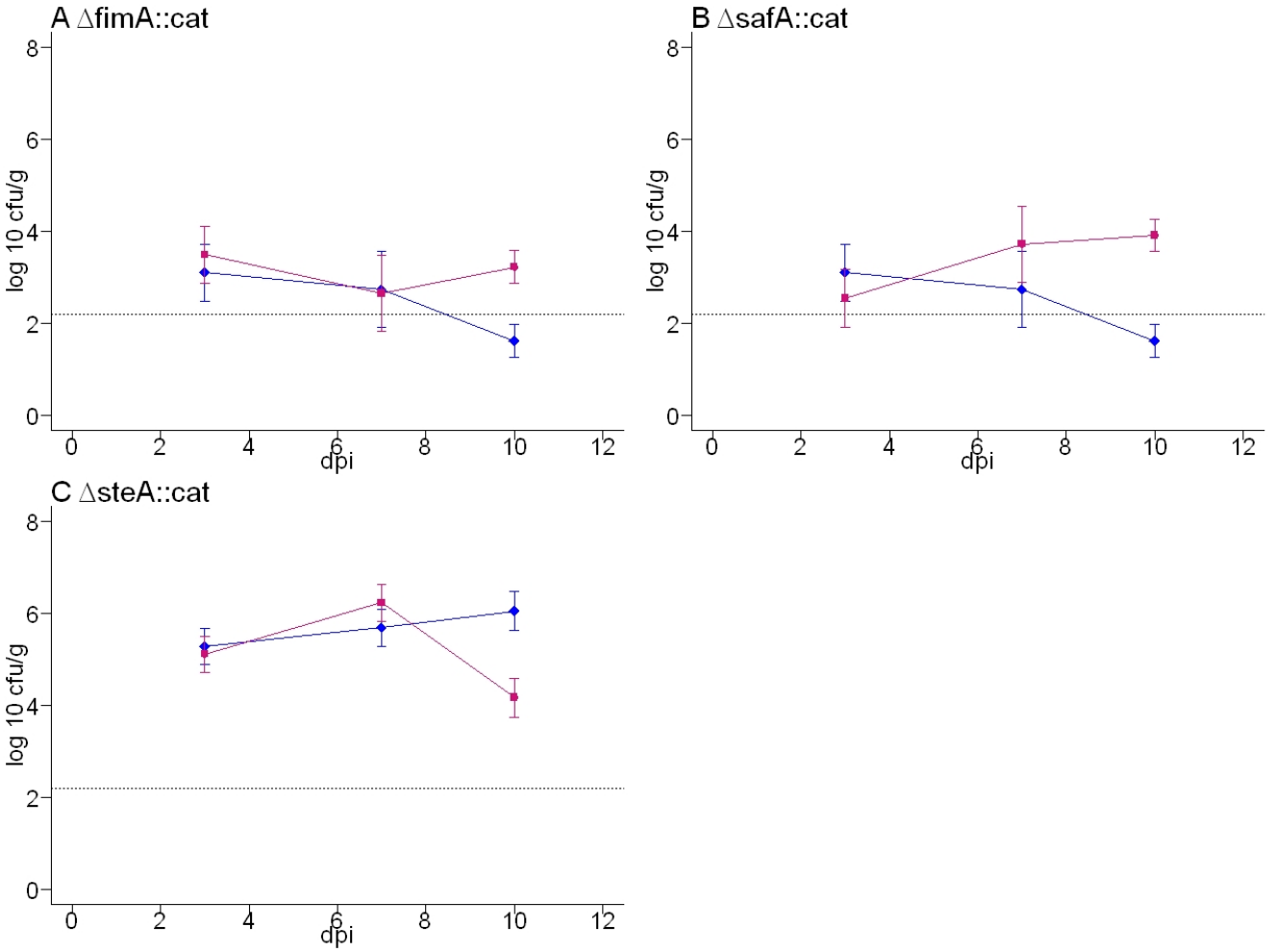


Figure 3
Total caecal load of *S*. Enteritidis S1400nal^R wild-type, Δ*fimA*::*cat*, Δ*steA*::*cat* and Δ*safA*::*cat* mutant strains at 3, 7 and 10 days post-oral inoculation of 18-day-old out-bred Rhode Island Red chickens. Blue lines with diamonds denote the wild-type strain and pink lines with squares denote the fimbrial mutants. The dashed line indicates the theoretical limit of detection by direct plating (2.2 $\log_{10}$ CFU/g). Samples positive only by enrichment culture were given an arbitrary value of 1 $\log_{10}$ CFU/g since at least one viable organism must have been present. The data reflect the mean ± standard error of the mean from five birds at each time interval. F-tests of the difference in recovery of wild-type and mutant strains at each time interval were performed and *P* values < 0.05 are marked with an asterisk.

tidis P125109 Δ*pegA* mutant resulted in total caecal counts that were not significantly different to the Δ*pegA* fimbrial mutant at both 1 and 3 days post-oral inoculation of chickens (P = 0.24 and P = 0.07, respectively). However, recoveries of the p$pegA_{rev}$-bearing strain were lower than for the Δ*pegA*::*cat* mutant alone at both time points, indicating that plasmid carriage may exert a slight fitness cost. The Δ*pegA* mutant harbouring p$pegA_{rev}$ was significantly attenuated compared to the parent strain at 1 and 3 days post-inoculation (P values < 0.0001). In contrast, introduction of the pACYC177-derived plasmid containing *pegA* in the sense orientation into the Δ*pegA* mutant partially restored the ability of the mutant strain to colonise the caeca at both time points relative to the wild-type strain (P = 0.0005 and P = 0.02 at 1 and 3 days post-inoculation, respectively) and to the Δ*pegA* fimbrial mutant (P = 0.0014 and P = 0.0005 at 1 and 3 days post-inoculation, respectively). Plating of tissue homogenates to media with or without ampicillin indicated that the plasmid was stably maintained in the absence of antibiotic selection *in vivo*. Taken together these data confirm that *pegA* plays a role in caecal colonisation of the avian intestines by *S*. Enteritidis.

Discussion

S. Enteritidis phage type 4 is an important zoonotic pathogen and the factors mediating persistence in the avian reservoir are ill-defined. Toward an understanding of the

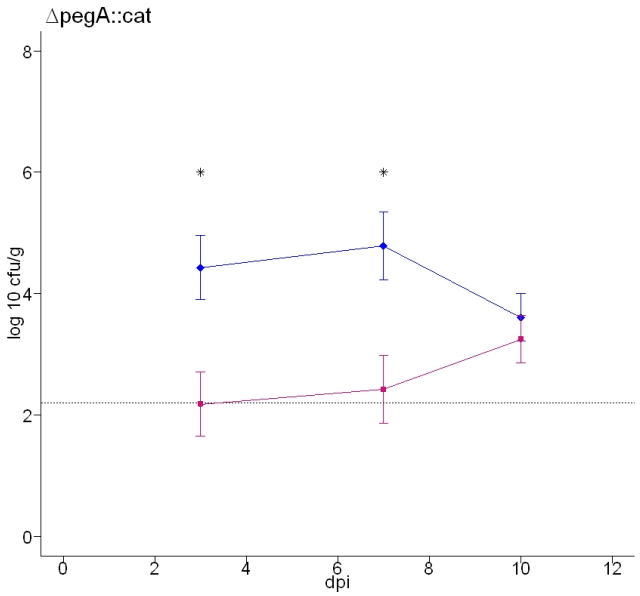


Figure 4
Total caecal load of *S*. Enteritidis S1400nal^R wild-type and Δ*pegA*::*cat* fimbrial mutant strains at 3, 7 and 10 days post-oral inoculation of 18-day-old out-bred Rhode Island Red chickens. Blue lines with diamonds denote the wild-type strain and pink lines with squares denote the fimbrial mutant. The dashed line indicates the theoretical limit of detection by direct plating (2.2 $\log_{10}$ CFU/g). The data reflect the mean ± standard error of the mean from five birds at each time interval. F-tests of the difference in recovery of wild-type and mutant strains at each time interval were performed and *P* values < 0.05 are marked with an asterisk.

molecular mechanisms by which *S*. Enteritidis colonises the chicken gut, the role of fimbriae was examined as these influence the carriage, virulence and tropism of other members of the Enterobacteriaceae. From the raw genome sequence of *S*. Enteritidis P125109, 13 intact chromosomal fimbrial loci were predicted. By comparing the sequence and distribution of the fimbrial loci with the published genomes of other *S. enterica* serovars *in silico*, no single locus correlated with host specificity. Microarray studies have indicated that a remarkable degree of conservation of fimbrial gene content exists among 26 *S*. Enteritidis isolates from varied geographical locations, hosts and years [55] and between strains of other broad host range serovars [56,57]. However, sequencing of such loci is required to determine if subtle differences in gene function exist.

Systematic mutagenesis of each major fimbrial subunit gene and screening in a chicken model revealed that the majority of major fimbrial subunits played no significant role in colonisation of the caeca (*P* values greater than 0.05). The absence of roles for *S*. Enteritidis Fim, Csg, Lpf and Sef fimbriae confirms previous reports that mutants lacking these fimbriae singly or in combination exhibit no significant defect in colonisation of chicks [19,20,22]. Conversely, the present study supports a role for Stb fimbriae in colonisation of the avian intestines by *Salmonella* that was suggested by the isolation of an *S*. Typhimurium *stbC* mutant by screening a library of signature-tagged mutants for attenuation in chicks [12]. The same screen of random mutants also identified attenuating mutations in *sthB* and *csgD*, however roles for *sthA* and *csgA* were not observed herein and studies with defined non-polar mutants and *trans*-complemented strains will be required to establish if the *sth* and *csg* loci play a conserved or serovar-specific role in colonisation of chickens. Owing to the relatively short-term nature of the studies reported here, we cannot preclude the possibility that the fimbrial subunits examined may play a role in longer-term persistence in the avian intestines or indeed tropism for the reproductive tract and egg, and further studies will be required to investigate this.

For the first time, we have shown that *S*. Enteritidis P125109 and S1400nal^R mutants of the novel Peg fimbrial operon show statistically significant attenuation in chickens that can be partially restored by plasmid mediated *trans*-complementation. A mutant in which the polar effects of the deletion of *pegA* are not predicted at the transcriptional or translational level was also attenuated; further implying that the phenotype of *pegA* insertion mutants is not due to altered expression of downstream genes. The inability of the p$pegA_{fwd}$ plasmid to fully restore colonisation to wild-type levels may reflect differences in the expression level of the fimbrial subunit *in vivo* and/or the fitness cost of maintaining the plasmid since recoveries of the Δ*pegA*::*cat* mutant bearing *pegA* on pACYC177 in the antisense orientation were slightly lower than for the mutant alone.

Assays with cultured cells did not indicate any significant role for *pegA* in adherence to primary chick kidney cells, HD11 avian macrophage-like cells or HEp-2 human laryngeal epithelial cells (data not shown) and there was no correlation between *in vitro* and *in vivo* results, regardless of the fimbriae examined. However, this is not unexpected as many fimbriae are known to be poorly expressed during culture in laboratory media [58], but are induced in bovine and murine intestinal lumen [58,59] and serve as antigens in mice [59].

Although there is attenuation of the *S*. Enteritidis *pegA* mutant, the *pegC* gene encoding a putative chaperone is a pseudogene in the sequenced strain of the poultry-adapted serovar *S*. Gallinarum, which implies that the possession of the entire fimbrial operon is unlikely to be a prerequisite for chicken colonisation. It should be noted that the tissue distribution of *S*. Enteritidis and *S*. Galli-

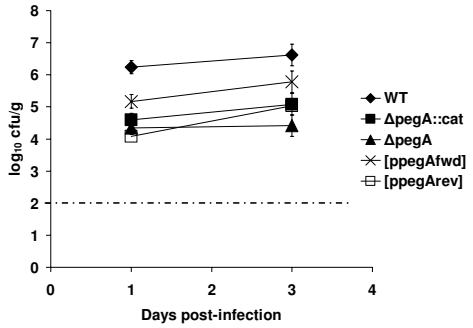


Figure 5
Plasmid-mediated *trans*-complementation of the Δ*pegA*::*cat* mutant of *S*. Enteritidis P125109 at 1 and 3 days post-oral inoculation of 18-day-old out-bred Rhode Island Red chickens. Total caecal load of the wild-type and mutant strain were compared to those of the P125109 Δ*pegA*::*cat* insertion mutant and Δ*pegA* strains in which *pegA* was introduced on plasmid pACYC177 in either the forward or reverse orientation relative to the promoter of the kanamycin resistance gene. The data represent the mean total caecal load ± standard error of the mean from five birds at each time interval for each strain. The dashed line indicates the theoretical limit of detection by direct plating (2.2 $\log_{10}$ CFU/g). F-tests of the difference in recovery of wild-type and mutant strains at each time interval were performed and *P* values < 0.05 are marked with an asterisk.

narum is markedly different in age-matched healthy out-bred birds, with *S*. Gallinarum causing severe systemic disease with little enteric involvement whereas *S*. Enteritidis colonises the gut to a high level [7].

The absence of significant roles for *S*. Enteritidis fimbrial loci in isolation may reflect redundancy or the existence of compensatory mechanisms, whereby the loss of single fimbriae may modulate expression of other fimbriae or colonisation factors. In a murine model deletions in the *S*. Typhimurium *lpf*, *pef*, *fim* and *csg* operons only moderately impaired virulence when tested individually, whereas a mutant containing all four deletions exhibited a 26-fold increase in the median lethal dose and reduced ability to colonise the intestinal lumen [15]. Further studies with *S*. Enteritidis strains harbouring multiple fimbrial mutations may be warranted. Transcriptome analysis of the expression of fimbrial genes in the mutant strains described herein may indicate whether cross-talk and compensation mechanisms exist, provided probes are used that discriminate between fimbrial transcripts in the absence of cross-hybridisation.

Conclusion

S. Enteritidis phage type 4 possesses thirteen chromosomally-encoded fimbrial loci, from which the predicted major fimbrial subunits of the majority can be deleted without significantly impairing caecal colonisation of chickens. Our data support the involvement of Stb fimbriae, previously suggested by screening of signature-tagged mutants of *S*. Typhimurium in poultry, and reveal for the first time that PegA influences caecal colonisation of chickens by *S*. Enteritidis. Since StbA and PegA serve as antigens in mice and vaccination with a cocktail of purified fimbrial subunits is partially protective in a murine model [59], further studies are required to evaluate the efficacy of subunit or live-attenuated vaccines that exploit the data presented here for control of zoonotic *S. enterica* serovars in poultry.

Authors' contributions

DJC annotated and mutated the fimbrial loci, characterised each mutant *in vivo* and drafted the manuscript. AJB, SDH, AMB and VLD provided valuable assistance to the chicken colonisation studies. NRT and MW supported bioinformatic analysis of the fimbrial operons and co-supervised DJC. PAB originally conceived the study. MPS led study design, data interpretation and manuscript revisions. EM and MJ provided a supervisory role.

Additional material

Additional file 1

Organisation of the fimbrial operons of* S. *Enteritidis P125109. *The image shows the gene organisation of each of the fimbrial operons, the Pfam domains within the fimbrial operons and the %GC content.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-8-228-S1.doc]

Additional file 2

Conservation of the nucleotide sequences of* S. *Enteritidis strain P125019 genes across sequenced strains of other* S. enterica *serovars. *The table provides the percent nucleotide identity of each fimbrial gene in several serovars of* Salmonella *compared with S. Enteritidis P125109.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-8-228-S2.doc]

Additional file 3

3a. Dam methylase target sequence GATC within and proximal to* S. *Enteritidis P125109 fimbrial operons. *3b. Putative homo-polymeric tracts in the S. Enteritidis P125109 genome sequence. The tables indicate regions of potential phase variable targets.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-8-228-S3.doc]

Additional file 4

S. ***Enteritidis P125109 colonisation of Rhode Island Red Chickens at 1, 3 and 7 days post-infection.*** *The graph shows the colonisation of S. Enteritidis P125109 at different organs within the chicken, at different times post-infection.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2180-8-228-S4.pdf]

Acknowledgements

The authors gratefully acknowledge funding via a CASE studentship from the Biotechnology & Biological Sciences Research Council and Lohmann Animal Health Ltd., and the support of the Department for the Environment, Food & Rural Affairs (grants OZ0319 and OZ0320). The authors wish to thank staff at the Experimental Animal House, IAH Compton, for technical assistance. Sequencing of the *S.* Enteritidis P125109 genome was supported via the Wellcome Trust Beowulf Genomics Initiative.

References

1. Voetsch AC, van Gilder TJ, Angulo FJ, Farley MM, Shallow S, Marcus R, Cieslak PR, Deneen VC, Tauxe RV, Emerging Infections Program FoodNet Working Group: **FoodNet estimate of the burden of illness caused by non-typhoid *Salmonella* infections in the United States.** *Clin Infect Dis* 2004, **38:**S127-134.
2. Braden CR: ***Salmonella enterica* serotype Enteritidis and eggs: a national epidemic in the United States.** *Clin Infect Dis* 2006, **43:**512-517.
3. Marcus R, Varma JK, Medus C, Boothe EJ, Anderson BJ, Crume T, Fullerton KE, Moore MR, White PL, Lyszkowicz E, Voetsch AC, Angulo FJ: **Emerging Infections Program FoodNet Working Group. Re-assessment of risk factors for sporadic *Salmonella* serotype Enteritidis infections: a case-control study in five FoodNet Sites, 2002–2003.** *Epidemiol Infect* 2007, **35:**84-92.
4. Altekruse SF, Bauer N, Chanlongbutra A, DeSagun R, Naugle A, Schlosser W, Umholtz R, White P: ***Salmonella enteritidis* in broiler chickens, United States, 2000–2005.** *Emerg Infect Dis* 2006, **12:**1848-1852.
5. Altekruse S, Koehler J, Hickman-Brenner F, Tauxe RV, Ferris K: **A comparison of *Salmonella enteritidis* phage types from egg-associated outbreaks and implicated laying flocks.** *Epidemiol Infect* 1993, **110:**17-22.
6. Patrick ME, Adcock PM, Gomez TM, Altekruse SF, Holland BH, Tauxe RV, Swerdlow DL: ***Salmonella enteritidis* infections, United States, 1985–1999.** *Emerg Infect Dis* 2004, **10:**1.
7. Wallis TS, Barrow PA: ***Salmonella* epidemiology and pathogenesis in food-producing animals.** In *Escherichia coli and Salmonella: Cellular and Molecular Biology [online]* Edited by: Dougan G. Washington DC: ASM Press; Module 8.6.2.1.
8. Hensel M, Shea JE, Gleeson C, Jones MD, Dalton E, Holden DW: **Simultaneous identification of bacterial virulence genes by negative selection.** *Science* 1995, **269:**400-403.
9. Shea JE, Hensel M, Gleeson C, Holden DW: **Identification of a virulence locus encoding a second type III secretion system in *Salmonella Typhimurium*.** *Proc Natl Acad Sci USA* 1996, **93:**2593-2597.
10. Turner AK, Lovell MA, Hulme SD, Zhang-barber L, Barrow PA: **Identification of *Salmonella* Typhimurium genes required for colonisation of the chicken alimentary tract and for virulence in newly hatched chicks.** *Infect Immun* 1998, **66:**2099-2106.
11. Tsolis RM, Townsend SM, Miao EA, Miller SI, Ficht TA, Adams LG, Baumler AJ: **Identification of a putative *Salmonella enterica* serotype Typhimurium host range factor with homology to IpaH and YopM by signature-tagged mutagenesis.** *Infect Immun* 1999, **67:**6385-6393.
12. Morgan E, Campbell JD, Rowe SC, Bispham J, Stevens MP, Bowen AJ, Barrow PA, Maskell DJ, Wallis TS: **Identification of host-specific colonisation factors of *Salmonella enterica* serovar Typhimurium.** *Mol Microbiol* 2004, **54:**994-1010.
13. Lawley TD, Chan K, Thompson LJ, Kim CC, Govoni GR, Monack DM: **Genome-wide screen for *Salmonella* genes required for long-term systemic infection of the mouse.** *PLoS Pathogens* 2006, **2:**e11.
14. Carnell SC, Bowen A, Morgan E, Maskell DJ, Wallis TS, Stevens MP: **Role in virulence and protective efficacy in pigs of *Salmonella enterica* serovar Typhimurium secreted components identified by signature-tagged mutagenesis.** *Microbiol* 2007, **153:**1940-1952.
15. Velden AW Van der, Baumler AJ, Tsolis RM, Heffron F: **Multiple fimbrial adhesins are required for full virulence of *Salmonella Typhimurium* in mice.** *Infect Immun* 1998, **66:**2803-2808.
16. Edwards RA, Schifferli DM, Maloy SR: **A role for *Salmonella* fimbriae in intraperitoneal infections.** *Proc Natl Acad Sci USA* 2000, **97:**1258-1262.
17. Weening EH, Barker JD, Laarakker MC, Humphries AD, Tsolis RM, Baumler AJ: **The *Salmonella enterica* serotype Typhimurium *lpf, bcf, stb, stc, std*, and *sth* fimbrial operons are required for intestinal persistence in mice.** *Infect Immun* 2005, **73:**3358-3366.
18. Ledeboer NA, Frye JG, McClelland M, Jones BD: ***Salmonella enterica* serovar Typhimurium requires the Lpf, Pef, and Tafi fimbriae for biofilm formation on HEp-2 tissue culture cells and chicken intestinal epithelium.** *Infect Immun* 2006, **74:**3156-3169.
19. Allen-Vercoe E, Woodward MJ: **Colonisation of the chicken caecum by afimbriate and aflagellate derivatives of *Salmonella enterica* serotype Enteritidis.** *Vet Microbiol* 1999, **69:**265-275.
20. Allen-Vercoe E, Sayers AR, Woodward MJ: **Virulence of *Salmonella enterica* serotype Enteritidis aflagellate and afimbriate mutants in a day-old chick model.** *Epidemiol Infect* 1999, **122:**395-402.
21. Allen-Vercoe E, Woodward MJ: **The role of flagella, but not fimbriae, in the adherence of *Salmonella enterica* serotype Enteritidis to chick gut explant.** *J Med Microbiol* 1999, **48:**771-780.
22. Thorns CJ, Turcotte C, Gemmell CG, Woodward MJ: **Studies into the role of the SEF14 fimbrial antigen in the pathogenesis of *Salmonella enteritidis*.** *Microb Pathog* 1996, **20:**235-246.
23. Rajashekara G, Munir S, Alexeyev MF, Halvorson DA, Wells CL, Nagaraja KV: **Pathogenic role of SEF14, SEF17, and SEF21 fimbriae in *Salmonella enterica* serovar Enteritidis infection of chickens.** *Appl Environ Microbiol* 2000, **66:**1759-1763.
24. De Buck J, van Immerseel F, Haesebrouck F, Ducatelle R: **Effect of type 1 fimbriae of *Salmonella enterica* serotype Enteritidis on bacteraemia and reproductive tract infection in laying hens.** *Avian Pathol* 2004, **33:**314-320.
25. Cogan TA, Jørgensen F, Lappin-Scott HM, Benson CE, Woodward MJ, Humphrey TJ: **Flagella and curli fimbriae are important for the growth of *Salmonella enterica* serovars in hen eggs.** *Microbiol* 2004, **150:**1063-1071.
26. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, Quail MA, Stevens M, Jones MA, Watson M, Barron A, Layton A, Pickard D, Kingsley RA, Bignell A, Clark L, Harris B, Ormond D, Abdellah Z, Brooks K, Cherevach I, Chillingworth T, Woodward J, Norberczak H, Lord A, Arrowsmith C, Jagels K, Moule S, Mungall K, Saunders M, Whitehead S, Chabalgoity JA, Maskell D, Humphreys T, Roberts M, Barrow PA, Dougan G, Parkhill J: **Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways.** *Genome Res* 2008, **18:**1624-1637.
27. Nuccio SP, Baumler AJ: **Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek.** *Microbiol Mol Biol Rev* 2007, **71:**551-575.
28. Datsenko KA, Wanner BL: **One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products.** *Proc Natl Acad Sci USA* 2000, **97:**6640-6645.
29. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R, Wilson RK: **Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2.** *Nature* 2001, **413:**852-856.
30. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N,

Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG: **Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18.** *Nature* 2001, **413:**848-852.

31. Deng W, Liou SR, Plunkett G 3rd, Mayhew GF, Rose DJ, Burland V, Kodoyianni V, Schwartz DC, Blattner FR: **Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18.** *J Bacteriol* 2003, **185:**2330-2337.
32. Chiu CH, Tang P, Chu C, Hu S, Bao Q, Yu J, Chou YY, Wang HS, Lee YS: **The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen.** *Nucl Acids Res* 2005, **33:**1690-1698.
33. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J: **ACT: the Artemis Comparison Tool.** *Bioinformatics* 2005, **21:**3422-3423.
34. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945.
35. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucl Acids Res* 1999, **27:**573-580.
36. Cherepanov PP, Wackernagel W: **Gene disruption in *Escherichia coli*: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant.** *Gene* 1995, **158:**9-14.
37. Chang AC, Cohen SN: **Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid.** *J Bacteriol* 1978, **134:**1141-1156.
38. Nurmi E, Rantala M: **New aspects of *Salmonella* infection in broiler production.** *Nature* 1973, **241:**210-211.
39. Hammar M, Bian Z, Normark S: **Nucleator-dependent intercellular assembly of adhesive curli organelles in *Escherichia coli*.** *Proc Natl Acad Sci USA* 1996, **93:**6562-6566.
40. Salih O, Remaut H, Waksman G, Orlova EV: **Structural analysis of the saf pilus by electron microscopy and image processing.** *J Mol Biol* 2008, **379:**174-187.
41. Abraham JM, Freitag CS, Clements JR, Eisenstein BI: **An invertible element of DNA controls phase variation of type I fimbriae of *Escherichia coli*.** *Proc Natl Acad Sci USA* 1985, **82:**5724-5727.
42. Blyn LB, Braaten BA, Low DA: **Regulation of *pap* pilin phase variation by a mechanism involving differential *dam* methylation states.** *EMBO J* 1990, **9:**4045-4054.
43. Meyer TF, van Putten JP: **Genetic mechanisms and biological implications of phase variation in pathogenic neisseriae.** *Clin Microbiol Rev* 1989, **2:**S139-145.
44. Old DC, Corneil I, Gibson LF, Thomson AD, Duguid JP: **Fimbriation, pellicle formation and the amount of growth of salmonellas in broth.** *J Gen Microbiol* 1968, **51:**1-16.
45. Old DC, Duguid JP: **Selective outgrowth of fimbriate bacteria in static liquid medium.** *J Bacteriol* 1970, **103:**447-445.
46. Swenson DL, Clegg S: **Identification of ancillary *fim* genes affecting *fimA* expression in *Salmonella typhimurium*.** *J Bacteriol* 1992, **174:**7697-7704.
47. Norris TL, Kingsley RA, Baumler AJ: **Expression and transcriptional control of the *Salmonella typhimurium lpf* fimbrial operon by phase variation.** *Mol Microbiol* 1998, **29:**311-320.
48. Nicholson B, Low D: **DNA methylation-dependent regulation of *pef* expression in *Salmonella typhimurium*.** *Mol Microbiol* 2000, **35:**728-742.
49. Balbontin R, Rowley G, Pucciarelli MG, Lopez-Garrido J, Wormstone Y, Lucchini S, Garcia-Del Portillo F, Hinton JC, Casadesus J: **DNA adenine methylation regulates virulence gene expression in *Salmonella enterica* serovar Typhimurium.** *J Bacteriol* 2006, **188:**8160-8168.
50. Chessa D, Winter MG, Nuccio SP, Tükel C, Bäumler AJ: **RosE represses Std fimbrial expression in *Salmonella enterica* serotype Typhimurium.** *Mol Microbiol* 2008, **68:**573-587.
51. Suar M, Jantsch J, Hapfelmeier S, Kremer M, Stallmach T, Barrow PA, Hardt W-D: **Virulence of broad- and narrow-host-range *Salmonella enterica* serovars in the streptomycin-pretreated mouse model.** *Infect Immun* 2006, **74:**632-644.
52. Fanelli MJ, Sadler WW, Franti CE, Brownell JR: **Localization of *Salmonellae* within the intestinal tract of chickens.** *Avian Dis* 1971, **15:**366-375.
53. Barrow PA, Simpson JM, Lovell MA: **Intestinal colonisation in the chicken by food poisoning *Salmonella* serotypes; microbial characteristics associated with faecal excretion.** *Avian Pathol* 1988, **17:**571-588.
54. Knodler LA, Bestor A, Ma C, Hansen-Wester I, Hensel M, Vallance BA, Steele-Mortimer O: **Cloning vectors and fluorescent proteins can significantly inhibit *Salmonella enterica* virulence in both epithelial cells and macrophages: implications for bacterial pathogenesis studies.** *Infect Immun* 2005, **73:**7027-7031.
55. Porwollik S, Santiviago CA, Cheng P, Florea L, McClelland M: **Differences in gene content between *Salmonella enterica* serovar Enteritidis isolates and comparison to closely related serovars Gallinarum and Dublin.** *J Bacteriol* 2005, **187:**6545-6555.
56. Porwollik S, Boyd EF, Choy C, Cheng P, Florea L, Proctor E, McClelland M: **Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays.** *J Bacteriol* 2004, **186:**5883-5898.
57. Anjum MF, Marooney C, Fookes M, Baker S, Dougan G, Ivens A, Woodward MJ: **Identification of core and variable components of the *Salmonella enterica* subspecies I genome by microarray.** *Infect Immun* 2005, **73:**7894-7905.
58. Humphries AD, Raffatellu M, Winter S, Weening EH, Kingsley RA, Droleskey R, Zhang S, Figueiredo J, Khare S, Nunes J, Adams LG, Tsolis RM, Bäumler AJ: **The use of flow cytometry to detect expression of subunits encoded by 11 *Salmonella enterica* serotype Typhimurium fimbrial operons.** *Mol Microbiol* 2003, **48:**1357-1376.
59. Humphries A, Deridder S, Bäumler AJ: ***Salmonella enterica* serotype Typhimurium fimbrial proteins serve as antigens during infection of mice.** *Infect Immun* 2005, **73:**5329-5338.

INFECTION AND IMMUNITY, Oct. 2011, p. 4105–4121 Vol. 79, No. 10
0019-9567/11/$12.00 doi:10.1128/IAI.01390-10

Salmonella enterica Serovar Typhimurium Colonizing the Lumen of the Chicken Intestine Grows Slowly and Upregulates a Unique Set of Virulence and Metabolism Genes[▽]

P. C. Harvey, M. Watson,† S. Hulme,‡ M. A. Jones,‡ M. Lovell,‡ A. Berchieri, Jr.,§ J. Young, N. Bumstead, and P. Barrow*

Institute for Animal Health, Compton Laboratory, Compton, Newbury, Berkshire RG20 7NN, United Kingdom

Received 31 December 2010/Returned for modification 26 February 2011/Accepted 11 July 2011

The pattern of global gene expression in *Salmonella enterica* serovar Typhimurium bacteria harvested from the chicken intestinal lumen (cecum) was compared with that of a late-log-phase LB broth culture using a whole-genome microarray. Levels of transcription, translation, and cell division *in vivo* were lower than those *in vitro*. *S.* Typhimurium appeared to be using carbon sources, such as propionate, 1,2-propanediol, and ethanolamine, in addition to melibiose and ascorbate, the latter possibly transformed to D-xylulose. Amino acid starvation appeared to be a factor during colonization. Bacteria in the lumen were non- or weakly motile and nonchemotactic but showed upregulation of a number of fimbrial and *Salmonella* pathogenicity island 3 (SPI-3) and 5 genes, suggesting a close physical association with the host during colonization. *S.* Typhimurium bacteria harvested from the cecal mucosa showed an expression profile similar to that of bacteria from the intestinal lumen, except that levels of transcription, translation, and cell division were higher and glucose may also have been used as a carbon source.

Salmonella enterica serovars Typhimurium and Enteritidis are the two *S. enterica* serovars most frequently associated with human food poisoning, with 1.4 million cases reported in the United States in 1999 (26) and an estimated 192,703 cases in the European Union in 2004 (4). Poultry and poultry products are generally considered to be major sources of human infection (3, 65). Healthy adult chickens generally show no clinical disease following oral infection with these serovars (6, 70). Infection of birds more than a few days old with *S.* Typhimurium or *S.* Enteritidis results in asymptomatic cecal colonization with persistent shedding of organisms, resulting in carcass contamination at slaughter and entry into the human food chain. The ecology of colonization of birds of this age is complex (21). In contrast, infection within a few hours of hatching, as can occur in hatcheries, when the chicken is immunologically immature and possesses a rudimentary gut flora, not only results in massive multiplication in the alimentary tract but can also result in severe systemic disease in the bird (6, 73).

Although intestinal colonization is central to entry into the human food chain, either through carcass contamination or by preceding systemic infection and subsequent egg contamination, the mechanism whereby *S. enterica* serovars colonize and interact with the host in the early stages of infection is still poorly understood. Screening of randomly generated mutant libraries of *S.* Typhimurium and more targeted studies have provided some insight into the bacterial genes required for colonization of chickens which are several weeks old and possess a gut flora. Type I and other fimbriae, including those encoded by the *stb*, *csg*, and *sth* operons (22, 31, 59), are thought to be involved in attachment of *Salmonella* and *Escherichia coli* bacteria to the mucosal layer or even to epithelial cells. Lipopolysaccharide is also thought to be involved, but it is unclear how (20, 59, 82). Additionally, global regulatory genes and a number of metabolic functions, including serine and citrate utilization, together with heat shock conditions, appear to contribute to the process in adult birds (59). Although some of the genes identified indicate that a close association with the gut mucosa is important in *Salmonella* colonization, the metabolic behavior of bacteria in the gut of newly hatched chickens is still poorly understood. Microbial behavior under these circumstances is very different from that in older birds. Viable numbers of *Salmonella* bacteria colonizing the cecum are much higher in younger than in older birds, and the interactions between the bacteria may more closely resemble those in stationary-phase broth cultures (100), where competition for nutrients under the prevailing redox conditions is at least known to be involved. Some studies also indicated the importance of proton-translocating proteins in colonization (44, 100; S. Muhammad, M. A. Jones, and P. Barrow, unpublished). Other factors, including some secreted proteins, contribute in different hosts, but it is again unclear how (52, 59, 82).

The numerical predominance of *Salmonella* bacteria in the ceca of young chicks following experimental infection allows

* Corresponding author. Mailing address: School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington, Loughborough, Leicestershire LE12 4RD, United Kingdom. Phone: 44 115 951 6428. Fax: 44 115 951 6415. E-mail: paul.barrow@nottingham.ac.uk.

† Present address: The Roslin Institute, The University of Edinburgh, Roslin Biocentre, Roslin, Midlothian EH25 9PS, Scotland, United Kingdom.

‡ Present address: School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington, Loughborough, Leicestershire LE12 5RD, United Kingdom.

§ Present address: Faculdade de Ciencas Agrárias e Veterinárias, Universidade Estadual Paulista, 14870-000 Jaboticabal, São Paulo, Brazil.

▽ Published ahead of print on 18 July 2011.

effective analysis of the bacteria in the absence of other organisms, and gene transcription pattern analysis at the genome level is thus possible. A whole-genome array derived from *S.* Typhimurium was used to investigate gene expression of the virulent avian phage type 14 strain *S*. Typhimurium F98 (70, 82, 100), harvested directly from chick ceca and compared with expression patterns from bacteria grown in broth *in vitro*. This approach, at least with *Campylobacter jejuni*, has demonstrated successfully that expression profiles under these conditions do resemble those observed in older, fully colonized birds (92).

MATERIALS AND METHODS

Chick colonization and sample collection. One hundred chickens from a brown-egg commercial laying line (Lohmann) were hatched in prefumigated incubators. Chickens were housed in fumigated cages and handled with sterile gloves to avoid contamination. One hundred chicks were infected orally within 12 h of hatching (to avoid the development of gut flora) by gavage with 0.1 ml of an *S.* Typhimurium F98 (70, 82, 100) culture grown for 16 h in LB broth at 37°C in a shaking incubator (150 rpm) and diluted to contain 10^7 CFU/ml. Only sterile water was provided, since the yolk sac is not fully resorbed for up to 3 to 4 days, providing sufficient food for the experimental period. At 16 h postinfection, the birds were killed individually and the cecal contents were removed immediately from the exposed ceca by syringe and mixed with Tri Reagent (Sigma). The cecal contents from seven of the birds were collected separately and stored on ice to be used for viable count estimations. The cecal contents from each group of birds, mixed with Tri Reagent, were pooled prior to extraction and purification. The purified RNA was further treated with DNase I and cleaned using RNeasy mini columns (Qiagen) and then concentrated further by RNA precipitation using 3 M sodium acetate. RNA was used only when the quality and concentration were optimal, as determined by spectrophotometer (Pharmacia). The experiment was repeated three times. Viable count estimations were made by plating decimal dilutions on MacConkey agar to allow the presence of any contaminating colonies among the predominant non-lactose-fermenting *Salmonella* bacteria to be detected. In the three experiments, the numbers of *Salmonella* bacteria were between 8.95 and 10.20 $\log_{10}$, and lactose fermenters or other colony types were not detected (<2 $\log_{10}$ per g).

Patterns of *in vivo* gene expression were compared with those of bacteria grown *in vitro*. For these controls, total RNA was extracted in the same way from three cultures of *S.* Typhimurium F98, in which 2 ml of an overnight LB broth culture was inoculated into 200 ml of prewarmed LB broth and incubated with shaking (150 rpm) for 3 h at 37°C. Cultures were pretreated with RNA Protect (Qiagen) before being centrifuged at 5,000 × *g* for 10 min at 20°C prior to RNA extraction.

Harvesting of *Salmonella* from the mucosal wall. In addition to harvesting the cecal contents, material was taken from the cecal mucosa for analysis by microarray. Samples were extracted by emptying the ceca with gentle pressure and then opening the walls of the ceca lengthwise and shaking the cecal walls in RNA Protect (Qiagen) to release bacteria from the surface. RNA from the ceca and from the washings from the mucosal wall was isolated using standard cleanup procedures, and samples from the same experiments were pooled. RNA from both samples was amplified using a MessageAmp II-bacteria kit (Ambion) per the manufacturer's instructions. RNA quality and concentration were determined with a spectrophotometer (Pharmacia). Gene expression was compared with that of the luminal samples.

Microarray hybridization. The *S.* Typhimurium array was printed as described previously (25). Total bacterial RNA was isolated from chicken ceca and from *in vitro* cultures grown in LB broth. The DNase-treated total *in vivo*- and *in vitro*-grown RNA was converted to fluorescently labeled cDNA using indirect labeling techniques (2, 25). Briefly, 15 μg of the total RNA samples from chick cecal contents was reverse transcribed (SuperScript II; Invitrogen) in the presence of 1 μl of deoxynucleoside triphosphates (dNTPs) (2.5 mM concentration each of dATP, dCTP, and dGTP and 1 mM concentration of dTTP [Amersham]), 1.5 μl of aminoallyl-dUTP (Sigma), and 30 μg of pd(N_6) (Amersham) in a total volume of 12 μl. This mixture was incubated overnight at 42°C before the reaction was stopped, and the mixture was cleaned with 450 μl of water in triplicate using Microcon units (YM-30; Millipore).

Two cDNA probes were labeled with 100 mg of Cy3 (*in vivo* lumen sample) or Cy5 (*in vitro* or *in vivo* mucosal sample) (monofunctional dyes; Amersham). The Cy3- and Cy5-labeled probes were combined and cleaned using a QIAquick PCR purification kit (Qiagen). The probe was dried in a speed vacuum before it was resuspended in a total volume of 25 μl of hybridization buffer (3× SSC [1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate], 25 mM HEPES, yeast tRNA, 50× Denhardt's solution, 10% [wt/vol] SDS), heated for 2 min, and then cooled in the dark. The probe was applied directly to the array with a clean coverslip placed on top. The probe was hybridized for 16 h at 63°C in a humidified slide chamber (Telechem, Inc., CA). The slide was postprocessed as described previously (25). Slides were scanned using a commercial laser scanner (GenePix 4000A; Axon Instruments, MDS, Sunnyvale, CA).

Data analysis. Fluorescence intensities of the signal and background were calculated for each spot using image analysis software (GenePix Pro 3.0; Axon Instruments). Three biological replicates each of both the *in vitro*-grown RNA and the *in vivo*-harvested RNA were compared. The data were analyzed using the Limma package (71). The data were first normalized within arrays using the Loess method (72) and then normalized between arrays in order to scale the log ratios to have the same median absolute deviation (MAD) across arrays (95). A linear model was then fitted for each spot across the series of arrays. The resulting *P* values were adjusted according to the false-discovery-rate method of Benjamini and Hochberg (8). Functional annotations were linked to the genes from the NCBI file NC_003197.ptt (http://www.ncbi.nlm.nih.gov/nuccore/16763390).

RT-PCR. The data were validated by quantitative reverse transcriptase PCR (qRT-PCR) of 15 genes which were differently regulated in the lumen samples to confirm gene expression ratios (87). Primers (Table 1) and fluoroprobes were designed using Primer Express software (PE Applied Biosystems) and purchased from Sigma-Genosys Europe Ltd. (Cambridge, United Kingdom). One-step qRT-PCR was performed in triplicate by using a mix of 2 ng/μl DNase-treated total RNA, gene-specific primers (50 nM) and probes (100 nM), and reverse transcriptase qPCR master mix (RT-QPRT-032X; Eurogenetic, EGT Group, Belgium). The concentrations of primers and template in each reaction mixture were determined by construction of a standard curve, starting with 200 ng total RNA and 500 nM primer and using 10-fold dilutions from 10^{-1} to 10^{-5}. Three total RNA samples were analyzed in triplicate in PCRs, and three replicate values were used to generate the standard curves. Amplification and detection of specific primers were performed using the ABI Prism 7700 sequence detection system (PE Applied Biosystems, Warrington, United Kingdom). The cycle parameters were as follows: an initial cycle of 48°C for 30 min and 95°C for 10 min and then 40 cycles of 95°C for 15s and 60°C for 1 min. The results were expressed in terms of threshold cycle value, the cycle at which the change in the reporter dye passes a significant threshold value above background. The fold changes in gene expression calculated from the qRT-PCR data were converted to $\log_2$ values and plotted against the changes calculated from the array data, which had also been $\log_2$ converted.

Creation of mutants. Insertion mutants using kanamycin or streptomycin/spectinomycin resistance cassettes were prepared as single mutants using standard procedures detailed elsewhere (82, 83, 100). Briefly, oligonucleotide primers were used to amplify upstream and downstream fragments, which were then joined together by an additional overlap extension PCR using the same two fragments as a template. This allowed the introduction of a KpnI site in the middle of the combined fragment and an XhoI and BglII (or, in the case of the *cobS* and *cbiA* mutants, XbaI) site at each end. This construct was incorporated into the suicide vector pDM4 (54), and the Km^r GenBlock insertion was introduced into the KpnI site. Spectinomycin and streptomycin (Spc-Str) resistance insertions were made in the same way. The cassette was in pHP45ΩSpc (H. Krisch, Département de Biologie Moleculaire, Université de Genève, Switzerland). A single-base-pair change generated a BamHI site in the middle of the fragment that enabled an Spc-Str resistance cassette to be inserted after base 406 of the open reading frame (ORF), and XbaI sites were incorporated into each end of the fragment for cloning into pDM4. Oligonucleotide primers are shown in Table 1. These pDM4 derivatives were maintained in *E. coli* strain SM10λ*pir* (83) and were introduced into the recipient *Salmonella* strains by conjugation. Transconjugants were isolated on selective medium supplemented with either streptomycin or kanamycin (25 μg/ml), and their sensitivities to chloramphenicol were then tested to identify those that resulted from a recombinational double-crossover event that had not incorporated any pDM4 DNA. The mutation was transduced into a fresh culture using P22 HT *int* (5). Transductants were checked by PCR using primers from the 3′ end of the cassette and the 5′ end of the structural gene, which generated a single DNA fragment in each of the mutants but not in the parent strain.

Double mutants were prepared with the creation of the additional mutation in a single mutant background using the alternative resistance cassette.

Assessment of colonization ability. Colonization was checked in specific-pathogen-free (SPF) day-old Light Sussex chickens obtained from the Poultry

TABLE 1. Oligonucleotide primers used for mutant production

Gene	Oligonucleotide sequences of primers[a]	Enzyme	Resistance cassette
argA	TCA**CTCGAG**GCAAAGAGGTGTGCCGTG	XhoI	Km
	GCCGCTGGGCCGCTGG**GGTACC**ARGACGGCGTGG ATT	KpnI	
	CTCGCCTCGTGCCAT**GGTACC**CCAGCGGC	KpnI	
	CGC**AGATCT**TAACCCTAAATCCGCCATCA	BglII	
potG	TCA**CTCGAG**ACGAAAGTGAAGAGCGGA	XhoI	Km
	GATAAAAAGCTG**GGTACC**AGGATGCACCTTGAA	KpnI	
	CGACCACCGAGGCAT**GGTACC**CAGCTTTTTATC	KpnI	
	CGG**AGATCT**CCGTCGGCACCACACAGCTC	BglII	
csgA	TCA**CTCGAG**GGATCAAAACTATTGTCCGT	XhoI	Km
	AATGCTCA**GGTACC**GCCGTTATGATTACC	KpnI	
	ATAACGGC**GGTACC**TGAGCATTTATCAGT	KpnI	
	CGC**AGATCT**TAGCGCAGACGCTAAATTAA	BglII	
metF	CGT**CTCGAG**GACATGAAGAAAATTCAACT	XhoI	Km
	TTATTCCA**GGTACC**GCTCTTTGATGCCTT	KpnI	
	CAAAGAGC**GGTACC**TGGAATAACGGTATC	KpnI	
	TCC**AGATCT**TGGCAAATGGCATAACTCAT	BglII	
ttrB	TCA**CTCGAC**CGCTGATTCTCTGGAGGA	XhoI	Km
	CTTGTACCG**GGTACC**CCGGCACAC AGG	KpnI	
	GTCCCTGGATGCCTG**GGTACC**GGCCATAGG	KpnI	
	CGC**AGATCT**TGGCAATGTGGACGGGAG	BglII	
ttrS	TCA**CTCGAC**CCCGGCTTGTTGTTGATC	XhoI	Km
	ACTGGGCCG**GGTACC**CGTCCACCAGTC	KpnI	
	CCGCCTGAGCCGCAT**GGTACC**CGCCCCAGT	KpnI	
	GCG**AGATCT**TCATCCAGTAGATGAAT	BglII	
pduA	TCA**CTCGAC**CCATGCGAGGTCTTTATG	XhoI	Km
	CGCGGCGAT**GTTACC**CGGTCAAAG	KpnI	
	TGCATCGGTGGCCGC**GGTAAC**ATCGCCGCG	KpnI	
	CGC**AGATCT**CCACCAGCTGACTGCTGC	BglII	
eutR	TCA**CTCGAC**GAGAGCCTCCCCATCAAT	XhoI	Km
	GTGGCCAGC**GTTACC**TGCACAAAGCCC	KpnI	
	CTAGCGCTGGAGGTA**GTTACC**GCTGGCGAG	KpnI	
	CGG**AGATCT**GTCGGAGGGCCGGGCGTC	BglII	
btuB	TCA**CTCGAC**AAGCCTGCGGCATCCTCC	XhoI	Km
	CTCCGCTAT**GGTACC**TTCCGATGCTAT	KpnI	
	GCGCTTTGTAGGAGG**GGTACC**ATAGCGGAG	KpnI	
	CGG**AGATCT**CGGTGGGACGAGGTTCAG	BglII	
cobS	GA**TCTAGA**ACGAATCTGCTGTTTGCGCT	XhoI	Km
	CAGCAG**GGTACC**TAGCGGAATACCACACCAG	KpnI	
	CCGCTA**GGTACC**CTGCTGACCGGTGGTTTTCA	KpnI	
	AG**TCTAGA**ACAGAGCCAGCAGAAAGATC	BglII	
	CAGCAG**GGATCC**TAGCGGAATACCACACCAGG	BamH1	Spc
	CCGCTA**GGATCC**CTGCTGACCGGTGGTTTTCA	BamH1	
cbiA	CA**TCTAGA**AAGGCATCACGCATTTATTC	XhoI	Km
	CGTTAT**GGTACC**AATGGCATTTTTGAGGAGCT	KpnI	
	GCCATT**GGTACC**ATACGGTGATGTTAAAACAT	KpnI	
	TG**TCTAGA**CAGCCAGTGCTGCACCATTT	BglII	
	TAACATCACC**GGATCC**GCCGCCAG	BamH1	Spc
	AGCAATCATGGCATG**GGATCC**GGTGATGTT	BamH1	

[a] Underlining and boldface indicate enzyme sites.

Production Unit, Institute for Animal Health. Birds were maintained in cages at 33°C with water and received no food prior to oral inoculation.

Colonization ability was assessed in two ways. First (100), groups of 10 chickens were inoculated orally within 24 h of hatching with 0.1 ml of an undiluted broth culture of the strain (mutant or parent of nalidixic acid-resistant [Nal^r] *S.* Typhimurium F98) to be tested. They were then given access to a vegetable protein-based diet (SDS, Manea, Cambridgeshire, United Kingdom). Twenty-four hours later, 3 birds were killed and the numbers of bacteria of the inoculated strain in the ceca were enumerated. The remaining 7 birds were inoculated orally with 0.1 ml of a 1:1,000 dilution of a broth culture of an Spc^r mutant of the parent F98 strain. Three days later, all birds were killed and the numbers of bacteria of both strains in the cecal contents were counted on brilliant green containing either sodium nalidixate (20 μg/ml) and novobiocin (1 μg/ml) or spectinomycin (50 μg/ml) (Sigma).

Second, at 1 day of age, groups of 20 chickens were inoculated orally with 0.1 ml of an overnight LB broth culture of cecal contents obtained from healthy, adult SPF chickens to prevent the development of systemic disease. They were then given access to feed, as described above. Twenty-four hours later, the chickens were infected orally with 10^8 CFU of either a spontaneous Nal^r mutant of *S.* Typhimurium F98 or a Nal^r mutant with a single or double insertion mutation in selected genes in 0.1 ml of LB broth. At 1, 2, and 3 weeks after inoculation, cloacal swabs were taken from each bird and plated in a standard manner (6) on brilliant green agar containing sodium nalidixate (20 μg/ml) and novobiocin (1 μg/ml) to obtain a semiquantitative enumeration of the bacteria excreted.

Virulence assays. Selected mutants of *S.* Typhimurium F98 were tested for their virulence for newly hatched Rhode Island Red chickens. The mutations were transferred by P22 transduction (5) to strain 4/74, which is virulent for mice (SL1344) (83), for assessment of virulence in BALB/c mice. Virulence was assessed by oral inoculation of groups of 20 newly hatched chickens with 0.1 ml or of 10 BALB/c mice with 50 μl of a broth culture diluted to contain 10^6 CFU in this volume. Morbidity and mortality were recorded over a 3-week period. Signs in chickens included anorexia and a disinclination to drink, standing with head and wings lowered, and caked feces around the vent. Mice became unsteady and had a "starry" coat. These signs are generally predictive of severe disease and death, and animals with signs of disease were killed humanely. Animals showing signs typical of salmonellosis were killed humanely, and their livers were cultured on MacConkey agar. Differences in mortality were analyzed by a χ^2 test.

Microscopy of cecal contents. Eight newly hatched chickens were inoculated orally with 0.1 ml of a 1/1,000 dilution of an overnight LB broth culture of *S.* Typhimurium within 8 h of hatching. Eighteen hours later, all birds were killed and cecal contents were harvested into universal bottles and stored at 4°C. They were diluted 1:100 in phosphate-buffered saline (PBS) and observed within 1 to 2 h by phase microscopy. The number of bacterial cells that showed evidence of division, expressed as a proportion of the total, was counted for each sample. Bacteria which were attached or had a visible septum were regarded as in the process of division. Motility and general cell shape were also observed.

Microarray data accession numbers. Raw data have been deposited in GEO (http://www.ncbi.nlm.nih.gov/pubmed/11752295), platform GPL6439 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6439), and series GSE10337 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10337).

RESULTS

Transcription profile from within the cecal contents. RNA extracted from *S.* Typhimurium bacteria from the luminal contents of the ceca of day-old chicks was compared to that from the *in vitro* cultures. The genes were grouped by clusters of orthologous groups of proteins (COGs) classification and are shown in Fig. 1. This overarching classification indicated major changes resulting from adaptation to the cecal environment. Overall, 17% of the 4,457 *S.* Typhimurium coding sequences (CDS) present on the array showed changes in expression during infection. Of these, 282 CDS were upregulated more than 2-fold, including genes associated with amino acid, carbohydrate, coenzyme, and lipid transport. A total of 464 CDS were downregulated more than 2-fold, including genes associated with cell cycle regulation, translation, and DNA replication. Total RNA was extracted from five noninfected birds to determine if the cecal contents alone produced a cross-reaction with the array; no cross-reaction was detected (data not shown).

Genes which showed statistically significant differential expression between *in vivo* and *in vitro* conditions (2-fold change, $P < 0.05$) were considered to be of interest. The genes with increased and decreased levels of expression which fulfilled this criterion are listed in Tables 2 and 3, respectively.

Compared with *in vitro*-grown luminal bacteria, significant changes were observed in genes associated with the following factors. (i) Relating to cell division, 12 genes associated with

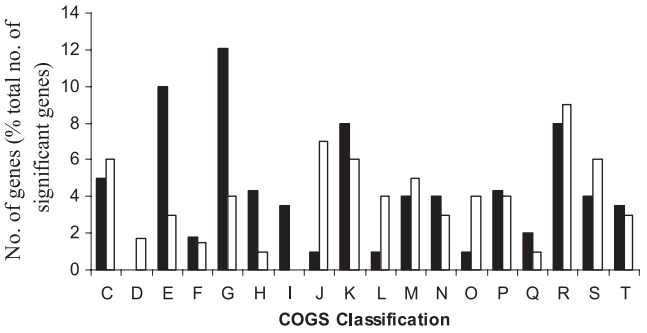


FIG. 1. Comparison of *S.* Typhimurium genes expressed in the lumens of newly hatched chicks with those expressed in *in vitro*-grown bacteria, classified according to COGs. Black bars, cecal contents; white bars, *in vitro*. The classified genes were found to be significantly different, with a >2-fold change in expression and a *P* value of less than 0.05. COGs classification abbreviations: C, energy production and conversion; D, cell cycle control, mitosis, and meiosis; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation; K, transcription; L, replication, recombination, and repair; M, cell wall/membrane biogenesis; N, cell motility; O, posttranslational modification, protein turnover, and chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport, and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms.

recombination, gene regulation, transcription, and chromosome replication, including *hupA*, *himA*, *ygiE*, and *dnaX*, were downregulated *in vivo*, compared to *in vitro*-grown bacteria. In addition, seven genes involved in cell division (including *ftsEKX*) were downregulated. There was a significant reduction in expression of 32 genes associated with translation, including *rplB* to *rplW*, *rpsAGJSP*, and *rpmBIJ*, following analysis of gene expression within the lumen of the cecum. Genes associated with DNA repair (including *dcm*, encoding DNA cytosine methylase, *recC*, and *sbcC*) were upregulated.

(ii) Regarding energy sources, the *prpBCDE* locus, but not *prpR*, its regulator, was significantly upregulated in the lumen. A number of genes in the *pdu* operon were upregulated, particularly the latter part, *pduK-pduV*. However, there was no associated upregulation of the *cob* or *cbi* genes. The *btuF* gene was found to be expressed, indicating utilization of an external source of cobalamin. Increased expression of genes in the *eut* (ethanolamine degradation) operon (*eutPQTDMN*) was detected. The low redox environment of the lumen is indicated by the significant upregulation of *ttrABC*, although *phs* and *asr* gene products were not significantly upregulated. Other genes associated with respiration with oxygen as the terminal electron acceptor, including *cydA*, *cyoCD*, *nuoEFIJ*, *frd*, and *napC*, were downregulated.

(iii) Regarding carbohydrates, a number of different loci involved in the utilization of carbohydrates showed different levels of up- and downregulation. Expression of *melA* was significantly upregulated in the lumen, although the changes in expression of *melB* and *melR* were not statistically significant. Four of the 11 genes (*yiaM*, *yiaN*, *lyxK*, *sgbH*) required for the catabolism of L-ascorbate to D-xylulose were upregulated. The gene encoding trehalose phosphate synthase, *otsA*, was also upregulated in the lumen, as were some unidentified genes,

TABLE 2. *S*. Typhimurium genes of interest which were upregulated during colonization of the cecal lumen, compared to gene expression in broth cultures[a]

COGs class	Locus tag	Gene	Function or product	Change in expression level (fold)	*P* value
Not in COGs	STM1144	*csgA*	Major curlin subunit precursor	3.9	0.002
	STM1601	*ugtL*	Putative exported protein	2.9	0.02
	STM1602	*sifB*	Secreted effector	3.4	0.01
	STM1384	*ttrC*	Tetrathionate reductase complex subunit C	2.3	0.03
	STM0550	*fimY*	Putative regulatory protein	3.1	0.01
	STM1143	*csgB*	Minor curlin subunit precursor	4.2	0.04
	STM3758	*fidL*	Putative inner protein	2.2	0.04
Amino acid transport and metabolism	STM0878	*potG*	Putrescine transporter	5.6	0.03
	STM2992	*argA*	*N*-α-acetylglutamate synthase	4.6	0.03
	STM4296	*adi*	Catabolic arginine decarboxylase	4.4	0.047
	STM1094	*pipD*	Pathogenicity island-encoded protein D	4.1	0.05
	STM4105	*metF*	5,10-Methylenetetrahydrofolate reductase	3.6	0.03
	STM3965	*metE*	5-Methyltetrahydropteriryltriglutamate-homocystein *S*-methyltransferase	2.2	0.04
	STM3086	*speA*	Arginine decarboxylase	3.6	0.02
	STM0887	*artJ*	Arginine transport system component	2.4	0.02
	STM2055	*pduU*	Polyhedral body protein	3.7	0.05
	STM2056	*pduV*	Propanediol utilization protein	2.8	0.01
	STM2469	*eutP*	Putative ethanolamine utilization protein	3.5	0.0006
	STM2468	*eutQ*	Putative ethanolamine utilization protein	3.0	0.007
	STM0877	*potF*	Putrescine transporter	2.5	0.05
	STM0878	*potG*	Putrescine transporter	5.6	0.03
Carbohydrate transport and metabolism	STM1928	*otsA*	Trehalose-6-phosphate synthase	4.3	0.03
	STM4298	*melA*	α-Galactosidase	3.2	0.04
	STM3674	*lyxK*	L-Xylulose kinase	2.0	0.047
	STM3675	*sgbH*	Putative 3-hexulose-6-phosphate isomerase	3.1	0.02
	STM0018		Putative exochitinase	3.0	0.014
	STM1560		Putative α-amylase	2.9	0.008
	STM3254		Putative fructose-1-phosphate kinase	2.7	0.009
	STM3671		Putative transporter	3.0	0.02
Energy production and conversion	STM0369	*prpC*	Putative citrate synthase	6.0	0.004
	STM1383	*ttrA*	Tetrathionate reductase complex subunit A	4.6	0.03
	STM2057	*pduW*	Propionate kinase	3.4	0.02
General function	STM0370	*prpD*	2-Methylcitrate dehydratase	7.1	0.001
Inorganic ion transport	STM2862	*sitB*	Putative ATP-binding protein	2.8	0.04
	STM0206	*btuF*	Putative periplasmic cobalamin-binding protein	2.6	0.01
	STM2863	*sitC*	Putative permease	2.3	0.05
Lipid	STM0371	*prpE*	Putative acetyl-CoA synthetase	3.5	0.003
Motility	STM0339	*stbB*	Putative fimbrial chaperone	2.8	0.01
	STM0195	*stfA*	Putative fimbrial subunit	3.4	0.02
	STM4593	*sthB*	Putative fimbrial usher protein	2.6	0.05
	STM0198	*stfE*	Putative minor fimbrial subunit	3.1	0.02
	STM0199	*stfF*	Putative minor fimbrial subunit	3.7	0.03
	STM0200	*stfG*	Putative minor fimbrial subunit	3.5	0.02
Replication	STM2150	*stcC*	Putative outer membrane protein	2.3	0.04
	STM0395	*sbcC*	ATP-dependent dsDNA exonuclease	2.9	0.01
	STM1992	*dcm*	DNA cytosine methylase	3.4	0.01
	STM2996	*recC*	Exonuclease V subunit	2.3	0.02
Secondary metabolites biosynthesis, transport and catabolism	STM2046	*pduK*	Polyhedral body protein	4.1	0.03
	STM2054	*pduT*	Polyhedral body protein	2.6	0.04
	STM2047	*pduL*	Propanediol utilization protein	3.5	0.03
	STM2465	*eutM*	Putative detox protein	2.4	0.02
	STM2464	*eutN*	Putative detox protein	2.4	0.02
Transcription	STM3964	*metR*	*metE*/*metH* regulator	2.2	0.03
	STM3756	*rmbA*	Putative cytoplsmic protein	3.1	0.04
	STM0552	*fimW*	Putative fimbrial protein	3.0	0.02
Translation	STM1909	*argS*	Arginine tRNA synthetase	2.0	0.04
Function unknown	STM1088	*pipB*	Secreted effector protein	5.5	0.01
	STM3764	*mgtC*	$Mg2^+$ transport protein	2.5	0.05
	STM0884		Putative inner membrane protein	4.8	0.04

[a] Genes selected as genes of interest showed a >2-fold increase in expression levels and a *P* value of <0.05.

TABLE 3. *S*. Typhimurium genes of interest which were downregulated during colonization of the cecal lumen, compared with expression in broth cultures[a]

COGs class	Locus tag	Gene	Function or product	Change in expression level (fold)	*P* value
Not in COGs	STM2770	*fljA*	Phase 1 flagellin repressor	2.5	0.03
	STM2304	*pmrD*	Polymyxin resistance protein B	6.02	0.005
Amino acid transport and metabolism	STM	*dsdA*	D-Serine deaminase	6.65	0.0036
	STM3244	*tdcB*	Threonine dehydratase	4.9	0.01
	STM3240	*tdcG*	L-Serine deaminase	2.6	0.03
Carbohydrate transport and metabolism	STM2433	*crr*	Glucose-specific IIA component	10.78	0.026
	STM4231	*lamB*	Maltoporin precursor	4.4	0.04
	STM2190	*mglB*	Galactose transport protein	5.6	0.001
	STM0684	*nagB*	Glucosamine-6-phosphate deaminase	2.5	0.05
	STM0685	*nagE*	*N*-Acetylglucosamine-specific enzyme IIABC	2.1	0.04
	STM2431	*ptsH*	Phosphohistidinoprotein-hexose phosphotransferase	9.00	0.014
Cell cycle	STM3569	*ftsX*	Putative cell division protein	3.3	0.008
	STN3570	*ftsE*	Putative cell division ATPase	3.3	0.03
	STM0960	*ftsK*	Cell division protein	3.5	0.02
Energy production and conversion	STM2320	*nuoJ*	NADH dehydrogenase I chain J	2.3	0.03
	STM2321	*nuoI*	NADH dehydrogenase I chain I	2.4	0.03
	STM2255	*napC*	Periplasmic nitrate reductase	3.0	0.02
	STM0440	*cyoD*	Cytochrome o ubiquinol oxidase subunit IV	3.4	0.04
	STM4340	*frdD*	Fumarate reductase membrane anchor polypeptide	3.6	0.05
	STM2325	*nuoE*	NADH dehydrogenase I chain E	4.1	0.03
	STM2324	*nuoF*	NADH dehydrogenase I chain F	3.5	0.05
	STM0441	*cyoC*	Cytochrome o ubiquinol oxidase subunit III	6.96	0.013
	STM0740	*cydA*	Cytochrome d terminal oxidase polypeptide subunit I	6.33	0.0062
General function	STM4361	*hfq*	Host factor I	10.08	0.0005
	STM1751	*hns*	DNA-binding protein HLP-II	8.73	0.0033
Cell motility	STM1959	*fliC*	Flagellin	14.89	0.00055
	STM1171	*flgN*	Putative FlgK/FlgL export chaperone	7.79	0.007
	STM1920	*cheW*	Chemotaxis docking protein	7.26	0.0027
	STM1183	*flgK*	Flagellar hook-associated protein 1	2.3	0.02
	STM4533	*tsr*	Methyl-accepting chemotaxis protein	3.9	0.03
	STM1915	*cheZ*	Chemotactic response protein	4.0	0.02
	STM2771	*fljB*	Phase 2 flagellin	4.0	0.01
	STM1921	*cheA*	Chemotaxis sensory histidine protein kinase	4.0	0.02
	STM1174	*flgB*	Flagellar basal body rod protein	4.2	0.05
	STM3577	*tcp*	Methyl-accepting transmembrane citrate/phenol chemoreceptor	5.2	0.002
Replication, recombination, and repair	STM1339	*himA*	Integration host factor alpha subunit	3.9	0.013
	STM3185	*yqiE*	ADP-ribose pyrophosphatase	3.6	0.007
	STM0484	*dnaX*	DNA polymerase III tau/gamma subunits	3.1	0.01
	STM4170	*hupA*	DNA-binding protein HU-alpha	6.36	0.00079
Signal transduction mechanisms	STM1916	*cheY*	Chemotaxis regulator	6.62	0.01
Transcription	STM0900		Putative helicase	12.5	0.002
	STM2875	*hilD*	Invasion protein regulatory protein	10.3	0.008
	STM1172	*flgM*	Anti-FliA factor	7.71	0.011
	STM2867	*hilC*	Invasion regulatory protein	5.89	0.014
	STM3245	*tdcA*	Transcriptional activator	3.0	0.005
	STM1956	*fliA*	Sigma 28	3.0	0.025
Translation	STM3728	*rpmB*	50S ribosomal subunit protein L28	5.92	0.02
	STM3440	*rplC*	50S ribosomal subunit protein L3	2.1	0.04
	STM3437	*rplB*	50S ribosomal subunit protein L2	2.22	0.02
	STM3425	*rplF*	50S ribosomal subunit protein L6	2.6	0.02
	STM3414	*rplQ*	50S ribosomal subunit protein L17	3.3	0.01
	STM3438	*rplW*	50S ribosomal subunit protein L23	2.9	0.01

Continued on following page

TABLE 3—*Continued*

COGs class	Locus tag	Gene	Function or product	Change in expression level (fold)	*P* value
	STM3430	*rplN*	50S ribosomal subunit protein L14	3.2	0.008
	STM3422	*rplP*	50S ribosomal subunit protein L16	3.3	0.01
	STM4393	*rpsR*	30S ribosomal subunit protein S18	2.2	0.03
	STM3441	*rpsJ*	30S ribosomal subunit protein S10	2.5	0.02
	STM0981	*rpsA*	30S ribosomal subunit protein S1	3.2	0.03
	STM3447	*rpsG*	30S ribosomal subunit protein S7	3.6	0.03
	STM3448	*rpsL*	30S ribosomal subunit protein S12	3.7	0.045
	STM3436	*rpsS*	30S ribosomal subunit protein S19	3.9	0.012
	STM3419	*rpmJ*	50S ribosomal subunit protein X	3.6	0.02
	STM1335	*rpmI*	50S ribosomal subunit protein L35	3.9	0.01
Translocation	STM3321	*yhbH*	Putative sigma N modulation factor	16.11	0.0049
Function unknown	STM2390	*yfcZ*	Putative cytoplasmic protein	9.52	0.0051
	STM3995	*yihD*	Putative cytoplasmic protein	6.99	0.00041
	STM2697		Phage tail-like protein	6.87	0.003
	STM4088	*yiiU*	Putative cytoplasmic protein	6.26	0.0017

[a] Genes selected as genes of interest showed a >2-fold change in expression levels and a *P* value of <0.05.

including STM0018, STM1560, and STM3254, classified as having a role in carbohydrate utilization. Interestingly, there was significant downregulation in glucose utilization genes, including *crr* and *ptsH* and genes involved in *N*-acetylglucosamine utilization, such as *nagBE*. The genes *lamB* and *mglB* were also downregulated.

(iv) With respect to amino acid utilization, there was a significant level of upregulation of expression of *metE*, *metF*, and *metR*, and upregulation of *adiA*, *speA*, *argA*, and *argS* indicated that arginine was being utilized by bacteria in the lumen. Interestingly, there was a significant upregulation in the expression of the *potFGHI* operon (putrescine transport) within the lumen. *tdcAB*, the transcriptional activator, and *tdcB*, involved in threonine utilization, were downregulated. In addition, *dsdA* and *tdcG*, involved in serine utilization, were also downregulated.

(v) For the bacterial surface, the majority of genes involved in flagellum production were downregulated in the lumen, including *flgM*, *flgN*, *flgK*, *flgB*, *fliC*, *fljB*, and *fliA*. There was also significant downregulation of chemotaxis genes *cheAWZ*, *tcp*, and *tsr*. Several fimbrial genes were significantly upregulated, including *stfAEFG*, *stbB*, *stjB*, *stcC*, *sthB*, *csgA*, and *csgB*. Parts of the *fim* operon, *fimY* and *fimW*, were also upregulated.

(vi) With respect to virulence factors, a small number of genes from *Salmonella* pathogenicity island 1 (SPI-1) were significantly upregulated, including *sitBC*, *sipD*, and *spaS*. The *sit* genes encode a part of an ABC transport system for the uptake of iron into the periplasmic space, indicating a potential function in colonization (94, 101). Two genes from SPI-1, *hilC* and *hilD*, were significantly downregulated in expression. Both of these genes are involved in the transfer of environmental signals to the central virulence gene regulator, HilA (23). This result is surprising, given the predicted effect of downregulating HilA. No significant change was observed for *hilA* expression, and a number of genes in SPI-1 were upregulated. This strongly suggests that in the lumen of the gut, a number of different factors are acting on the regulation of HilA. Little change in expression was detected within SPI-2 and SPI-4. Within SPI-3, *mgtC*, *rmbA*, and *fidL* and the colonization-associated genes *shdA* and *misL*, and in SPI-5, *pipB*, were found to be upregulated in the lumen. While a role for *mgtC* has been described for growth in low-magnesium environments (58), the requirement for *pipB*, *rmbA*, and *fidL* expression may represent redundant gene expression from the same island. There also seems no obvious reason for *pipB* expression to be required, as it is involved in intracellular kinesin binding (37).

Validation by quantitative RT-PCR. To validate the microarray results, RT-PCR was carried out on 15 selected genes showing different levels of expression within the lumen. The data for the 15 genes (Fig. 2) gave an r^2 value of 0.53, which was a good fit ($P = 0.0019$). The slope (2.27) indicated higher values by RT-PCR than by microarray.

Transcription profile from the mucosal wall. Patterns of gene expression in RNA extracted from *S.* Typhimurium bacteria from the washings of the cecal mucosa were compared to the data arising from the RNA harvested from the cecal lumen of day-old chicks. The genes were grouped by COGs classification and are shown in Fig. 3. A total of 33 genes were significantly (change of 2-fold, $P < 0.005$) upregulated at the

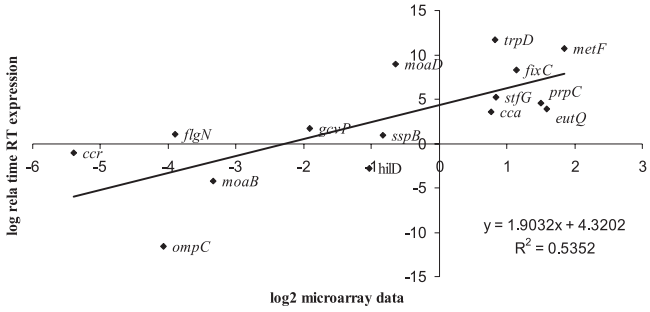


FIG. 2. Correlation between microarray and real-time RT-PCR expression values. Log_2-transformed expression values for 15 genes from total bacterial RNA extracted from day-old-chick cecal contents in triplicate. The best-fit linear regression line is shown together with the r^2 value and the calculated slope equation.

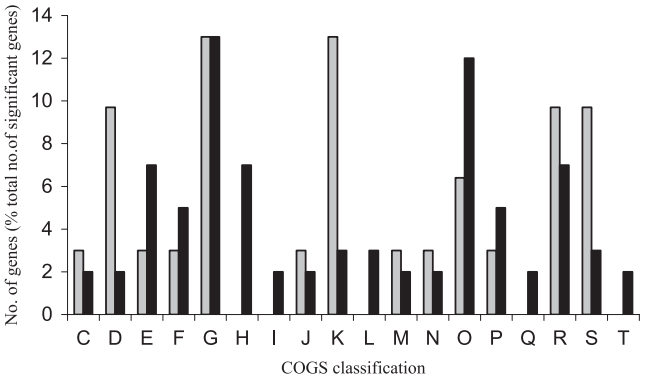


FIG. 3. Comparison of *S.* Typhimurium genes expressed at the mucosal wall with those expressed in the lumens of newly hatched chicks, classified according to COGs. Black bars, lumen; gray bars, mucosal wall. The classified genes were found to be significantly different, with a >2-fold change in expression and a *P* value of less than 0.05. COGs classification abbreviations: C, energy production and conversion; D, cell cycle control, mitosis, and meiosis; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation; K, transcription; L, replication, recombination, and repair; M, cell wall/membrane biogenesis; N, cell motility; O, posttranslational modification, protein turnover, and chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport, and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms.

mucosa, and 16 genes were significantly downregulated (Tables 4 and) 5.

Potentially significant changes in the mucosa, compared with luminal bacteria, were observed in genes associated with the following factors. (i) Relating to carbohydrate transport and metabolism, genes associated with glucose utilization were significantly upregulated at the mucosa, including *gmhA*, *ptsH*, and *crr*. A phosphotransferase suppressor of *ompF* was downregulated at the mucosal wall.

(ii) Regarding amino acid transport and metabolism, only one gene, *yhiP*, encoding a putative peptide transport protein, was significantly upregulated at the mucosa. However, three genes were significantly downregulated, including two encoding ABC transporter proteins and *carB*.

(iii) With respect to energy production and conversion, the cytochrome b_{562} gene, *cybC*, was significantly upregulated at the mucosal wall. Four genes, *pduDUW* and *ybhP*, plus a gene coding for a tetratricopeptide repeat protein were significantly downregulated.

(iv) With respect to cell division and transcription, three genes associated with cell division were upregulated, namely, *ftsA*, *yhdE*, and *mreB*. Four genes associated with transcription were also upregulated (including *rpoN* and *yciT*).

(v) Regarding translation and posttranslational modification, two genes, *tpx* and *sspA*, were significantly upregulated at the mucosal wall, as was *rpsM*, which encodes a 30S ribosomal subunit protein. However, only one gene, the ribosome stabilization factor gene *yfiA*, was downregulated significantly.

(vi) Interestingly, *pipB* was upregulated at the mucosal wall. *ugtL*, which provides resistance to antimicrobial peptides, was also upregulated.

It is appreciated that the number of genes showing changes in expression at the mucosa is very small compared with that of the bacteria in the lumen and that the overall patterns of expression in the two populations were almost identical.

Microscopy of bacteria from the mucosa. Phase-contrast microscopy of the cecal lumen and cecal mucosa from the above-described samples was used to estimate the numbers of dividing bacteria in these two sites in the cecum. Samples taken from eight chickens indicated that the percentage of dividing bacteria was higher at locations close to the mucosal wall than within the lumen (Fig. 4). There was no difference in bacterial cell size. No bacterial cells from the lumen showed evidence of motility (directional movement) in 10 microscopic fields, supporting the results of the gene expression studies above.

Colonization of chickens. Assessing the contribution to intestinal colonization of genes which were upregulated in the intestine was difficult in newly hatched chickens, since even serovars such as *Salmonella enterica* serovar Choleraesuis, which is unable to colonize the alimentary tract of adult birds, are nevertheless able to multiply in the guts of newly hatched chickens. We therefore decided to use a competition assay in which selected mutants are assessed for their ability to exclude a superinfecting parent strain inoculated 24 h later (100). This method is preferred to an assay in which both strains are inoculated simultaneously, because it allows an assessment of whether the mutant is utilizing the same nutrients under stationary-phase redox conditions as the parent strain which will compete with it. Our experience is that mutants which are sometimes noninhibitory in our assay are nevertheless frequently able to grow to equally high numbers as the parent strain when inoculated simultaneously (P. Barrow, M. Lovell, and M. A. Jones, unpublished results). Mutants were selected because the genes were relatively highly upregulated (*metF*, *csgA*, *argA*, and *potG*) or because they were linked metabolically (*ttrS*, *ttrB*, *pduA*, and *eut*). At the time of challenge, all mutants tested (*metF*, *csgA*, *ttrS*, *ttrB*, *pduA*, *eut*, *argA*, and *potG*) colonized the gut well, judging from the counts in the ceca of three birds killed at the time of challenge (Table 6). When the birds were killed 3 days after challenge, most mutants were still colonizing well, with the mean cecal count ranging from 7.03 to 7.71 log g^{-1}. Only the *argA*, *pduA*, *ttrS*, and *potG* mutants were found in low numbers. Despite this, all the mutants tested were able to exclude the parental challenge strains, with 6 of the 7 birds killed having challenge counts of <2 log, whereas the mean count of the challenge strain in birds which had not been previously inoculated with another strain was 5.11 (range, 4 to 7.20).

Because *pdu*, *ttr*, and *btu* genes were all upregulated in the intestine, and because these genes are all related to the anaerobic catabolism of 1,2-propanediol and ethanolamine, mutants with inactivated *ttr*, *pdu*, *eut*, or *btu* genes, and also *cob* and *cbi* operons, were tested for their ability to colonize the guts of 1-day-old chickens which had received gut flora preparations. The patterns of fecal excretion are shown in Fig. 5. The greatest reductions in fecal excretion from that of the parent strain were seen with the *pduA*, *ttrB*, and *cbiA* genes and the *cobS btuB* double mutant. Statistical significance was assessed using the χ^2 test. Statistically significant reductions in colonization were observed only with the *ttrB* mutant ($P < 0.01$). Additional

TABLE 4. *S.* Typhimurium genes upregulated at the mucosal wall[a]

COGs class	Locus tag	Gene	Function or product	Change in expression level (fold)	*P* value
Amino acid transport and metabolism	STM3592	*yhiP*	Putative peptide transport protein	2.08	0.049
Carbohydrate transport and metabolism	STM0310	*gmhA*	Phosphoheptose isomerase	2.06	0.037
	STM1558		Putative glycosyl hydrolase	2.13	0.017
	STM2431	*ptsH*	Phosphohistidinoprotein-hexose phosphotransferase	2.29	0.017
	STM2433	*crr*	Glucose-specific IIA component	2.4	0.036
Cell cycle control, mitosis, and meiosis	STM0132	*ftsA*	ATP-binding cell division protein	2.37	0.019
	STM3371	*yhdE*	Putative inhibitor of septum formation	2.26	0.0198
	STM3374	*mreB*	Rod shape-determining protein	2.13	0.014
Cell motility	STM1177	*flgE*	Flagellar hook protein	2.64	0.03
Cell wall/membrane biogenesis	STM0124	*murF*	D-Alanine-D-alanine ligase	2.09	0.044
Energy production and conversion	STM4439	*cybC*	Cytochrome b562	2.06	0.03
Function unknown	STM0119	*yabB*	Putative cytoplasmic protein	2.28	0.042
	STM1088	*pipB*	Secreted effector protein	2.35	0.016
	STM3347	*yhcB*	Putative periplasmic protein	2.42	0.037
General function prediction only	STM1581	*yddE*	Putative phenazine biosynthetic protein	2.02	0.045
	STM2580	*era*	GTPase	2.09	0.028
	STM2969	*ygdH*	Putative nucleotide binding	3.18	0.012
Inorganic ion transport and metabolism	STM4324	*cutA*	Putative periplasmic divalent cation tolerance protein	2.3	0.02
Not in COGs	STM0471	*ylaC*	Putative inner membrane protein	2.06	0.003
	STM1059	*ycbW*	Putative cytoplasmic protein	2.71	0.03
	STM1092		Putative cytoplasmic protein	2.32	0.024
	STM1601	*ugtL*	Putative membrane protein	2.1	0.04
	STM2983	*orfX*	Putative lipoprotein	2.09	0.01
Nucleotide transport and metabolism	STM1163	*pyrC*	Dihydroorotase	2.09	0.013
Posttranslational modification, protein turnover, and chaperones	STM1682	*tpx*	Thiol peroxidase	2.79	0.038
	STM3342	*sspA*	Stringent starvation protein A	2.23	0.041
Transcription	STM1704		Putative regulatory protein	2.73	0.022
	STM3320	*rpoN*	Sigma 54	2.09	0.022
	STM3515	*yciT*	Transcriptional activator	2	0.026
	STM4318		Putative acetyltransferase	2.198	0.019
Translation	STM3418	*rpsM*	30S ribosomal subunit protein S13	2.73	0.026

[a] Genes selected as genes of interest showed a >2-fold increase and a *P* value of <0.05.

reductions which were less significant were observed with *pduA* ($P = 0.03$) and *cobS* ($P = 0.1$) mutants.

None of these mutations produced any significant attenuation in the virulence of *S.* Typhimurium for mice or newly hatched chickens. Signs of severe systemic disease were observed in 8 to 10 of the 10 inoculated mice and in 15 to 20 of the 20 inoculated chickens, regardless of whether the strain was the virulent parent or a mutant strain ($P = 0.25$). Pure, heavy growth of *Salmonella* was obtained by culturing the livers of animals which were killed humanely.

DISCUSSION

The results here demonstrate that extensive transcriptional changes occur following infection of day-old chicks with *S.* Typhimurium, with many genes being downregulated in expression, indicating decreased metabolic activity from that of the broth culture. Those genes which were upregulated reflect a degree of adaptation to the luminal environment.

To study gene expression in *Salmonella* during colonization of chickens, the most appropriate model is generally regarded

TABLE 5. *S.* Typhimurium genes downregulated at the mucosal wall[a]

COGs class	Locus tag	Gene name	Function or product	Change in level of expression (fold)	*P* value
Amino acid transport and metabolism	STM0067	*carB*	Carbamoyl-phosphate synthase large subunit; putative ABC transporter	2.9	0.035
	STM1255		Periplasmic binding protein	2.49	0.044
	STM2055	*pduU*	Polyhedral body protein	2.097	0.031
	STM1257		Putative ABC transporter protein	2.03	0.006
	STM3594	*prlC*	Oligopeptidase A	2.09	0.0496
Carbohydrate transport and metabolism	STM3784		Putative phosphotransferase system mannitol/fructose-specific IIA domain	2.32	0.0199
Cell wall/membrane biogenesis	STM2120	*asmA*	Suppressor of OmpF assembly mutants	2.12	0.049
Energy production and conversion	STM2057	*pduW*	Propionate kinase	2.1	0.006
	PSLT027	*ccdA*	Antidote	2.14	0.045
	STM0813	*ybhP*	Putative cytoplasmic protein	2.09	0.025
	STM2007		Tetratricopeptide repeat protein	2.33	0.034
Not in COGs	STM0903		Putative chaperone	2.79	0.003
	STM2041	*pduD*	Propanediol dehydratase medium subunit	2.58	0.04
	STM3688		Putative cytoplasmic protein	2.16	0.0499
Replication, recombination, and repair	STM4168	*nfi*	Endonuclease V	2.22	0.011
Transcription	STM3773		Putative transcriptional regulator	2.07	0.019
Translation	STM2665	*yfiA*	Ribosome stabilization factor	2.75	0.01

[a] Genes selected as genes of interest showed a >2-fold change in expression levels and a *P* value of <0.05.

to be animals that are 2 to 6 weeks old and that have established gut floras which would be more dominant numerically than the colonizing pathogen. The constraints imposed by studying gene expression by microarray meant that experiments had to be performed in newly hatched chickens to avoid false-positive signals from the presence of numerically dominant flora components, such as *E. coli*. This model reflects the situation that occurs during infection in newly hatched chickens which does take place within hatcheries. Despite the shortcomings of this approach, the patterns of expression were closer to our preconceptions than we imagined. Similarly, patterns of global gene transcription in *Campylobacter jejuni* in a similar model were found to resemble those in older birds with gut floras (92), and other similar models (e.g., a streptomycin-treated mouse) have been used with *E. coli* with success (15, 44, 45).

The requirement for a large number of chickens to generate sufficient RNA also meant that bacteria present in the ceca of different birds would also likely have been present at different stages of the growth cycle, depending on whether the ceca were full, had just emptied, or were freshly filled (P. Barrow, un-

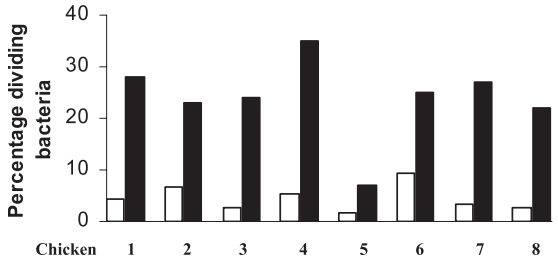


FIG. 4. Percentage of bacteria showing evidence of cell division out of the total number of bacteria observed by phase microscopy from the cecal lumens (white bars) or cecal mucosae (black bars) of 8 chickens infected orally with *S.* Typhimurium when less than 24 h old and killed 24 h later.

TABLE 6. Viable counts of test (Nal^r) and challenge (Spc) mutants of *S.* Typhimurium F98 in the ceca of newly hatched chickens in a competition assay[a]

Mutation	Viable count of[b]:		
	Test strain		Challenge strain postmortem
	At time of challenge	Postmortem	
metF	8.32, 8.42, 8.59	7.17 (6.60–7.82)	<2 (<2–<2)
csgA	8.40, 7.64, 8.04	7.79 (7.18–8.18)	<2 (<2–<2)
ttrS	7.73, 6.30, 7.08	6.95 (6.00–7.79)	<2 (<2–2.85)
ttrB	7.93, 8.43, 9.11	7.74 (7.00–8.00)	<2 (<2–<2)
pdu	8.08, 8.04, 8.00	6.60 (6.00–7.65)	<2 (<2–5.2)
eut	8.2, 8.88, 9.00	7.59 (7.18–7.68)	<2 (<2–<2)
argA	7.11, 7.88, 7.28	6.00 (6.00–6.90)	<2 (<2–2.78)
potG	7.90, 7.83	6.95 (6.70–7.28)	<2 (<2–<2)
Parent	8.80, 8.18, 8.28	7.00 (6.60–8.57)	<2 (<2–<2)
None	<2, <2, <2	<2 (<2–<2)	4.3 (4.00–7.82)

[a] Ten chickens were inoculated with the test strain. Three chickens were killed to enumerate this strain 24 h later at the time of challenge. All chickens were killed 3 days later to enumerate both strains in the ceca.

[b] Viable counts at time of challenge are presented for all three chickens. Viable counts postmortem of all other chickens are the means and ranges.

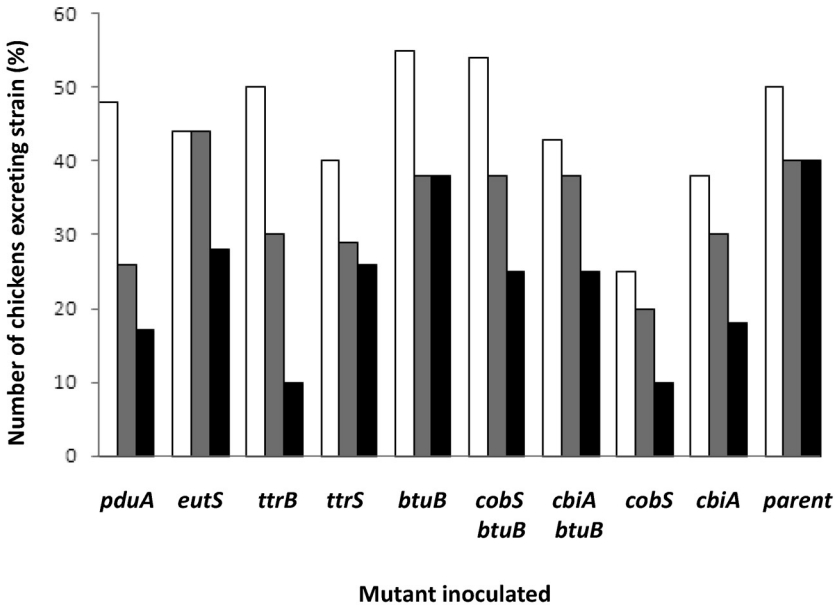


FIG. 5. Numbers of chickens, expressed as a percentage, which were excreting the inoculated mutant of a nalidixic acid-resistant derivative of *S.* Typhimurium F98 at 1 week (white bars), 2 weeks (gray bars), and 3 weeks (black bars) after oral inoculation of chickens possessing gut floras.

published). This potential variation had implications for measuring the expression of genes associated with logarithmic versus stationary-phase growth, but it did not appear to have a profound effect, judging from the patterns of expression observed.

Although we measured luminal gene expression, we were aware that the ceca contained heterogeneous environments, as indicated by the differing rates of cell division in the lumen and close to the mucosa. This was supported by differences in expression levels in genes associated with cell division, transcription, and translation. What was perhaps more surprising was that so few other genes were affected. The metabolic and virulence profiles from the bacteria harvested from the lumen fairly well reflected those from the mucosa, where most of the cell multiplication was taking place. This was reassuring. The small number of changes in expression at the mucosa from that of the luminal contents suggested that the two populations were very similar, but it offers some insight into the lifestyle of the bacteria close to the mucosa. The involvement of genes (*crr* and *gmhA*) in the uptake and metabolism of glucose, galactose, or mannose suggests that sialic acid from host cell membranes would likely act as a potential carbon source for bacteria close to the mucosa.

The data presented were also validated by the similar changes in expression observed in selected genes tested by RT-PCR, as has been found by other authors (2, 92).

The reduced level of cell division within the lumen, indicated by microscopy, together with the reduced expression of genes associated with cell division, transcription, and translation, suggests a greatly reduced rate of metabolism and growth at this site. There is a direct dependence of transcription and translation rates and gene doses on bacterial growth rates (49, 53), in addition to the dependence on total RNA quantity and ribosomal proteins (35, 47). The relationship between cell growth rates and expression of genes associated with cell division is less clear, since a key gene associated with the formation of the Z ring, *ftsZ*, is expressed independently of growth rate (88). The *ftsEKX* genes interact and form part of the divisome. Although functionality of several *fts* genes is not required for cell growth, as indicated by continued filamentous growth in *fts* mutants, *ftsE* mutants do show reduced growth rates, which can be suppressed at high osmolarities (18). Despite the apparent low rate of cell division and the shortage of nutrients in the lumen, there was evidence that propionate, 1,2-propanediol, and ethanolamine acted as important carbon sources. This was less apparent at the mucosal wall, where the *pduDUW* genes were significantly downregulated and the main source of nutrients was unclear, although there was some evidence for use of glucose in this niche but not in the lumen (see below). Interestingly, expression of SspA, stringent starvation protein A, was upregulated at the mucosal wall. This protein in *E. coli* was found to be induced during stationary phase and starvation for carbon, amino acids, nitrogen, and phosphate (34). It is thought to act as a global regulator. Its expression suggests that the bacteria were experiencing conditions of starvation, and though most bacterial multiplication is occurring close to the mucosa, this itself is not an ideal or static environment and it indicates the complexity of the environmental niches in the gut.

Within the lumen, degradation of 1,2-propanediol appeared to be occurring, although this generally requires endogenous adenosyl cobalamin (coenzyme B_{12}) biosynthesis. The *pdu* genes are contiguous and coregulated with cobalamin biosynthetic genes (*cob* and *cbi*) (49, 66). However, in the current experiments, there was no significant upregulation of the *cob* or *cbi* operons within the lumen. Some vitamin B_{12} is thought to be present in egg yolk (17) and would be present in the gut of newly hatched chickens, as the yolk sac is not fully resorbed for 3 to 4 days. This could be scavenged by BtuF (85), and BtuF in *Salmonella* is a periplasmic binding protein with a high affinity for vitamin B_{12} which was expressed within the lumen.

Genes involved in the catabolism of ethanolamine, which is derived in part from host cells and membranes, are encompassed in the *eut* operon (67, 75). The *eutT* gene encodes an adenosyltransferase, which is used to activate EutR, and in turn triggers transcription of the operon (68). Two of the genes, *eutM* and *eutN*, partially encode a metabolosome with the products of *eutSLK*, which encode the shell proteins of the metabolosome (74). The role of this structure was proposed to be to concentrate low levels of ethanolamine catabolic enzymes (10). The *eutD* gene encodes a phosphotransacetylase, which acts as a safety valve to minimize flux variations in a system which converts ethanolamine into acetyl coenzyme A (acetyl-CoA). The roles of *eutP* and *eutQ* remain unclear, though they were significantly upregulated in the cecal lumen.

Tetrathionate is one of the electron acceptors of choice for the utilization of ethanolamine and 1,2-propanediol (64) in the absence of oxygen. Other genes associated with respiration, including *cydA*, *cyoCD*, *nuoEFIL*, *frd*, and *napC*, were downregulated, suggesting that an anaerobic environment is present in the cecal lumen. This is in contrast to the findings of Jones et al. (44) showing that cytochrome *bd* oxidase was required for colonization of the streptomycin-treated mouse intestine by *E. coli*. These models are not strictly comparable, since the streptomycin-treated mouse will retain some gut flora, whereas there was virtually none in this series of experiments. In addition, we have found a degree of host specificity related to the likely route of respiration during intracellular *Salmonella* infection in chickens and mice (83) and the redox conditions implied therein. Tetrathionate is reduced to thiosulfide and further to H_2S with the products of *ttr*, *phs*, and *asr* genes. It is likely that the tetrathionate results, in part, from material from the yolk sac, which is rich in sulfur. The role of sulfur-based electron acceptors in respiration in the gut has been shown recently by Winter et al. (90), who demonstrated that in mice with acute intestinal infection, reactive oxygen is released, which generates thiosulfate to be used as an electron acceptor. The model used here involved birds in which, at the time of harvesting, no inflammation was visible. It seems likely that during a more established infection when inflammation and gut damage will also occur, similar events are likely to take place.

Expression of *ackA*, encoding acetate kinase, which balances acetate and acetyl coenzyme A production, and an alternative phosphate donor acetyl phosphate, was upregulated 2-fold. A significant role of substrate-level phosphorylation in chickens is further supported by the poor colonization ability of *ackA* and *pta* mutants (P. Barrow and M. A. Lovell, unpublished findings).

The results from *in vivo* studies with mutations affecting the complex interactions between propanediol and ethanolamine as carbon sources, tetrathionate as the electron acceptor, and cobalamin as a cofactor were ambiguous, probably indicating the degree of redundancy in these nutrients as carbon sources. Thus, although the *pduA* mutant, like the other mutants, was fully inhibitory in the competition assay, it colonized the gut less well in these birds and also colonized the birds with the floras less well, albeit with a reduction of marginal significance. The *eutS* mutant colonized the gut well, again indicating the degree of redundancy in carbon source availability in this complex niche. Thus, although genes may be upregulated, indicating metabolic activity, their mutation will divert metabolic activity to other catabolic pathways. Both the *ttrB* and *ttrS* mutants colonized less well in this assay, with only the *ttrB* being significantly reduced. The picture is confused by the fact that the double mutants with a *btuB* mutation colonized well, whereas the single *cobS* and *cbiA* mutants colonized less well, although not significantly so. The interaction between propanediol utilization with tetrathionate and with cobalamin is highly complex, and much of the nature of these interactions *in vivo* remains to be determined.

The breakdown of propionate occurs via the 2-methylcitrate cycle using the *prpBCDE* locus (33), encoding the propionate-degrading enzymes and carrying *prpR*, a transcriptional regulator (38) which was previously thought to act as a sensor for 2-methylcitrate, an intermediate of the breakdown pathway (60, 61, 81). Although *cobB* expression was also thought to be required (79), there was no significant difference between expression *in vivo* and *in vitro*. The absence of *cobB* expression may be compensated for by expression of *pduW*, which encodes propionyl coenzyme A, a precursor of 2-methylcitrate, and which was upregulated 4-fold. The *prpE* mutant showed no reduction in colonization ability from that of the parent strain (tested in a different assay; results not presented). However, given the other energy sources available to *Salmonella* within the lumen, this was not unexpected.

D-Glucose is taken up and concomitantly phosphorylated either by the glucose-specific enzyme II (EII) transporter or by the phosphoenol-pyruvate-dependent transporter (97). The phosphoryl group is transferred to glucose through enzyme I (encoded by *ptsI*) and the phosphohistidine carrier protein (encoded by *ptsH*) to sugar-specific EII, which consists of two subunits, *crr* and *ptsG*. At the mucosal wall, glucose may be a more important carbon source, with upregulation of *ptsH* and *crr*, though expression of *ptsI* and *ptsG* was not significant. However, in the cecal lumen, we think that a number of other carbohydrates may also have been utilized, most significantly melibiose and L-ascorbate, suggesting, with the downregulation of *crr* and *ptsH*, that glucose was not an available source. The breakdown of melibiose utilizes two genes, *melA* (α-galactosidase) and *melB* (transporter), and their expression is stimulated by MelR (77). Expression of *melA* was significantly upregulated in the lumen, although expression of *melB* and *melR* was not statistically significant. With the high levels of expression of *melA*, it suggests either that this compound may already have been present in the cell or that the product of the *melA* gene was being used to break down a second carbohydrate source.

Four of the 11 genes required for the catabolism of L-ascorbate to D-xylulose, which enters the pentose phosphate pathway, were upregulated. Generation of internal trehalose also appears to occur in the lumen, with the upregulation of *otsA*. This would fit with a model where the bacteria in the lumen are growing slowly or are under stress, as the trehalose operon is induced under these conditions in an RpoS-dependent manner (76, 84). Trehalose has been demonstrated to play a role in cell protection against stressful environmental conditions, such as osmotic stress and heat shock, and was proposed to have a role in survival but not virulence (39).

Several other sources of carbohydrates were not utilized in the lumen, including maltose and galactose, as indicated by

downregulation of *lamB* (30) and *mglB*, respectively. Again, this is in contrast to the findings of Jones et al. (45), which showed that maltose was important for *E. coli* colonization of the mouse intestine. These authors also found, in contrast to our previous findings, that glycogen was a significant carbon source (57). These results indicate the different responses in terms of gene expression and metabolism to colonizing different hosts, as recognized by Chang et al. (15) and as is found in those genes responsible for respiration during systemic *Salmonella* infection in chickens or mice (83).

Bacteria from the ceca demonstrated a requirement for methionine with significant levels of expression of *metE*, *metF*, with which *metH* forms the folate branch of the methionine pathway, and *metR*. The MetR protein acts as an activator for the transcription of *metE*, *metA*, *metF*, and *metH* (93). Homocysteine functions as a coregulator for MetR-mediated regulation and has a positive effect on the expression of *metE*, which encodes a transmethylase, and *metF*, which encodes 5,10-methylenetetrahydrofolate reductase but has a negative effect on *metA* and *metH*. The methylation of homocysteine, the final reaction prior to formation of methionine, is carried out via the vitamin B_{12}-independent enzyme MetE (94). Genes involved in the utilization of other amino acids, including threonine (*tdcB*) and serine (*dsdA* and *tdcG*), were downregulated in the cecal lumen. The downregulation of *tdcA*, the transcriptional activator of the *tdc* operon, suggests that there is little requirement for threonine or serine within the lumen. There was also a significant downregulation of genes involved in the biosynthesis of glycine and one-carbon units (*gcvH* and *gcvP*), suggesting that these amino acids were not essential for growth and survival in the lumen.

The environment within the chick cecum was thought to be very weakly acidic, at pH 6.5 to 7 depending on diet, in addition to being anaerobic. Several mechanisms of survival of *Salmonella* under acidic conditions have been well documented (27, 29), although it is fairly certain that this pH would not induce a strong acid tolerance response. Three acid-resistant (AR) mechanisms have been identified in *Escherichia coli*, including AR1, which involves RpoS and cyclic AMP (cAMP) enabling cells to resist a pH as low as 2.5 (28). The AR3 system involves an arginine decarboxylase and has recently been identified in *Salmonella* and expressed under anaerobic conditions (48). *S.* Typhimurium DT104 was found to induce an arginine-dependent AR response involving transcriptional activation of *adiA* and *adiC* genes by *adiY* (48). In the present study, expression of *adiA* and *speA* was detected at high levels *in vivo*, suggesting that *Salmonella* was degrading arginine to agamatine. Upregulation of the transcriptional regulator *adiY* and of *speB*, which converts agmatine to putrescine, was not detectable in the ceca. However, *Salmonella* was actively generating arginine, as indicated by the upregulation of *argA* and *argS*, and scavenging arginine, as indicated by expression of *artJ*, which encodes a binding protein for arginine. Interestingly, *Salmonella* expressed significant levels of genes from the *potFGHI* operon, which encodes an ATPase-binding, putrescine-specific uptake system. Polyamines have been found to increase survival in extremely acidic and other inimical environments (96). Whether these data indicate low pH at the microenvironmental level or resistance to another factor inimical to metabolism in a gross environment where the pH is close to neutral remains to be determined. Mutation of *argA* did not alter colonization ability or survival in day-old chicks, although a role for *speA* in the colonization of 2-week-old chickens was suggested by Morgan et al. (59).

Mutation of *potG* did not alter colonization of day-old chicks. Similarly, Morgan et al. (59) found that mutation of *potH* did not reduce colonization ability. This suggests that the *potFGHI* operon was not functioning to transport putrescine into the cell but may have been playing an alternative role.

As the evidence suggested an environment where oxygen concentrations were very low, a 3-fold increase in *dcm*, associated with DNA repair, was unexpected. Heithoff et al. (36) reported previously that *dam* mutants were virulent in mice, but the role of cytosine methylation (*dcm*) was unclear. It is important in the regulation of biological processes in plants and animals, but the role of *dam* in the methylation of adenine is more important. However, the results here suggest that in chicks, *dcm* may contribute to the survival of *Salmonella* within that environmental niche. Given the probably low oxygen content of the cecal lumen, suggested by the downregulation of *cydA*, *cyoCD*, *nuoEFIJ*, and *frd*, the expression of *recC* was unexpected. The protein encoded by this gene functions to repair damage to DNA caused by host-synthesized compounds. Mutations in *recA* and *recBC* were found to be highly sensitive to oxidative compounds synthesized by macrophages and avirulent in mice (11). Similarly, the expression of *sbcC* was unexpected. The protein encoded by this gene acts to restore recombination and to resist DNA damage. It suggests that radical oxygen molecules, which could be damaging to the chromosome, may exist within the lumen. Interestingly, expression of *tpx* was detected at the mucosal wall, suggesting a gradient of oxygen across the cecum itself. Bacteria protect themselves from reactive oxygen species with a range of antioxidant defense enzymes, including thiol peroxidase. It was found that *tpx* acts as a lipid peroxidase to inhibit bacterial membrane oxidation and acts as a principle antioxidant for *E. coli* during anaerobic growth (14). It is possible that *tpx* may be functioning in a similar way here. Again, the recent work by Winter et al. (90) is relevant here, since it indicates that the release of reactive oxygen species into the gut results from inflammation. Although there was no indication of any gross inflammatory response here, the induction of proinflammatory cytokines by invading bacteria is a rapid event (46) and begins to be apparent by 16 to 24 h postinfection of newly hatched chickens (91). This process would undoubtedly have started in the gut of the chickens examined here.

Bacteria in the lumen displayed poor motility compared to *in vitro*-grown bacteria, as demonstrated by phase-contrast microscopy. The lack of motility is further supported by the downregulation of a number of genes involved in flagellar structure and function in the cecal lumen. The majority of the genes involved in the process were downregulated in the lumen, including two regulatory genes, *flgM* and *flgN*, which act to regulate gene expression. *flgM* acts as anti-sigma factor 28, which binds sigma factor 28 until the completion of the hook-basal body unit (2). *flgN* has two roles (1): it acts as a sensor for late gene expression in flagellar assembly by promoting expression of *flgM* translation, and it is associated with hook-associated proteins to inhibit its translation on flagellar completion. The first hook-filament junction protein, encoded by *flgK*, was

downregulated, as was *flgB*, which forms part of the rod protein (95). Interestingly, *fliC* and *fljB*, which encode flagellin, were downregulated, as was *fljA*, which acts as a negative regulator for *fliC* expression (95). This suggests that no flagellin was produced in the chick lumen and, with the lack of expression of chemotaxis genes (*cheAWZ*, *tcp*, *tsr*), suggests that there is no major chemoattractant in the lumen which *Salmonella* bacteria would move toward. The downregulation in expression of *tcp* and *tsr* (41) suggests that neither citrate (*tcp*) nor serine (*tsr*) is present in the lumen. Stecher et al. (74) have shown that motility increases closer to the mucosa in the inflamed mouse intestine, although flagellation was less important in the noninflamed gut. We did not look at motility at the mucosa, but there would certainly not have been any gross inflammation during the short period of the experiments here.

Up to 13 different fimbrial operons have been suggested to be elaborated by *Salmonella* (40, 56). Some fimbrial genes are only expressed in particular environments (24). Within the chick lumen, several fimbrial genes were expressed, including *stfAEFG*, *stbB*, *stjB*, *stcC*, and *sthB*, suggesting that they may have a role in colonization or survival outside the host. The *stf* operon was found not to be essential for colonization by Clayton et al. (16). Morgan et al. (59) suggested that *stbC* and *sthB* contributed to colonization of older chickens. Genes required for biosynthesis of thin, curled fimbriae (*csgB* and *csgA*) were upregulated in the lumen as in macrophages (24). These are thought to have a role in adhesion, becoming associated with extracellular matrix, and are known to have a role in pathogenesis in *E. coli* (31). They appeared to play little role in our *in vivo* model.

The *fim* operon, encoding type 1 fimbriae, was downregulated in the lumen due to the upregulation in the expression of two regulatory genes, *fimY* and *fimW*. The role of the *fimY* gene in *S.* Typhimurium remains unclear, though it is essential for fimbrial production and acts as a coactivator with *fimZ* (79). *fimW* acts as a negative regulator and interacts with *fimZ*-mediated activation of *fimA* expression (78).

Salmonella pathogenicity islands (SPI) contain genes which confer virulence-associated functions upon the host bacterium, often mediated by secreted proteins. In *Salmonella*, many pathogenicity islands and other gene clusters have been well characterized, and expression of a number of genes from the 5 major islands has been detected. A small number of genes from SPI-1 were upregulated, including *sitBC*, which encodes an iron uptake system (101). The *sitABCD* operon is induced under iron-deficient conditions and is thought to play a role in iron acquisition in mice (42, 98). Interestingly, *hilC* and *hilD* are downregulated in the lumen. These genes encode transcriptional activators, which can bind to *hilA* and induce expression of three operons within SPI-1, namely, *inv-spa*, *prg-org*, and *sic-sip* (23). The high levels of repression of *hilD* suggest that expression of SPI-1 is inhibited, though expression of *sipD* and *spaS* was detected. *hilD* also plays a role in mediating the activities of SPI-1 and SPI-2 (12). The role of SPI-1 genes, and secreted proteins in general, in colonization in day-old chicks has not been widely investigated and is of considerable interest. Most SPI-1 genes were found not to be required for colonization of the cecal lumen of older birds by Morgan et al. (59), although they were required for colonization of the intestinal mucosa of calves. Recently, Jones et al. (43) found that SPI-1 did not play an essential role in systemic infection in 1-day-old birds. A subset of SPI-2 genes, including *ttrAC*, *ssaBCDMSU*, and *sseC*, were upregulated significantly in the lumen. Interestingly, gene expression was detected throughout SPI-2, though most of the changes were not significant. Regulation of expression of SPI-2 genes is thought to involve OmpR-EnvZ and PhoP-PhoQ (9), but none of the genes encoding these proteins showed significant alterations in their levels of expression. Again, Morgan et al. (59) found very few of these genes to be required for colonization of older birds. Interestingly, Wigley et al. (89) found in day-old chickens that SPI-1 contributed to and SPI-2 was essential for the virulence of *Salmonella enterica* serovar Pullorum in newly hatched chicks, where gut colonization does represent an early phase of the infection process in this infection model (73).

In SPI-3, *mgtC* was upregulated, along with *rmbA* and *fidL*. The *mgtC* gene forms a part of the *mgtBC* operon, which is positively regulated by magnesium, although the exact role of *mgtC* has not yet been clearly defined. This gene does not have a role in magnesium uptake, though it may have a role in long-term survival in macrophage cell lines (58), suggesting that *mgtC* may have a similar role here. Statistically nonsignificant increases in expression were also observed with *mgtA* and *mgtB*. No genes were expressed from SPI-4, and no role in colonization of chickens was observed by Morgan et al. (59). However, in SPI-5, *pipB*, whose role is unclear, was found to be upregulated in the lumen and at the mucosal wall. These authors also found that *pipB* contributed to colonization of older chickens (59). Although its role is unclear, it has a link with SPI-2 since this SPI is required for its secretion (50), though here those genes were not significantly upregulated.

It is worth noting that expression of *ugtL* was upregulated at the mucosal wall. This gene is required for resistance to the antimicrobial peptides maginin 2 and polymyxin B (69). Within the lumen, magnesium limitation appears to be occurring, but it is interesting that a defense peptide is being expressed by *Salmonella*, suggesting that a rapid response by the host to infection is occurring.

The exact role for secreted proteins in colonization is unclear. Older work (82) supported the hypothesis that cecal colonization, of chickens at least, by *Salmonella* was largely a physiological characteristic, since there seemed little ecological advantage in adhesion to the mucosa in an organ where the rate of flow of chyme was very low. However, the identification of some SPI genes in colonization (59, 82) suggested that colonization, as a virulence trait, might not be as straightforward as originally thought. More recent information indicated that a T cell-mediated response, rather than secretory antibodies, was central to immune clearance of *S.* Typhimurium from the chicken gut (7), suggesting that a close association with the mucosa may indeed be involved. The upregulation of fimbrial genes (59) supports this assertion.

The microscopic observations indicated a higher rate of cell division in the mucosa than in the lumen of the cecum. This was supported by the increased expression at the mucosa of *ftsA*, a gene involved in cell division which acts to anchor the protofilaments of bacterial tubulin, encoded by *ftsZ*, to the membrane (62). However, the presence of *ftsA* alone is not sufficient for the Z ring to form, and *zipA* (32) and other downstream genes required were not expressed at significant

levels at the mucosal wall. On the other hand, further support for active cell division occurring close to the mucosa comes from expression of *mreB*. The protein encoded by *mreB* has been found responsible for the rodlike shape of *Salmonella* bacteria (19). Recent work (86) suggests that the MreB protein directs the incorporation of new peptidoglycan into the wall, though the presence of *ftsZ* is required to direct the insertion. In addition, *yhdE*, which inhibits the formation of the septum, was expressed (12), suggesting that the cells were elongating after cell division at the point of sample collection. The MreB protein also contributes, it is thought, to chromosomal segregation (51).

These preliminary data suggest that *S.* Typhimurium bacteria in the cecal lumens of newly hatched chickens show downregulation of genes associated with transcription, translation, and cell division, all required for growth, whereas there was some expression of genes associated with cell division in bacteria harvested closer to the mucosa. It seems likely that concentrations of oxygen or of other electron acceptors and a variety of nutrients would be present closer to the mucosa than to the lumen, so these findings are not surprising. They are supported by earlier results with *E. coli* colonization of the mouse intestine, where there was evidence for most microbial growth taking place close to the mucosa (63), and our microscopic findings support this. The data suggest that several energy and carbohydrate sources are utilized which are different from those used in late-log-phase nutrient broth cultures, including propionate, ethanolamine, and 1,2-propanediol. Organisms in the lumen were poorly motile and showed a downregulation of genes associated with chemotaxis, though no genes associated with motility were identified or expressed at the mucosa. From our colonization studies, it was clear that several genes associated with propanediol catabolism under anaerobic conditions were involved in colonization, although the picture is obviously complex. However, in the light of previous findings where very few genes made great changes when tested as single mutations (16, 100), coupled with the degree of metabolic redundancy in enteric bacteria, this perhaps should not be surprising. This indicates that the colonization phenotype is a multifactorial characteristic with the main site of metabolic and other physiological activity close to the mucosa.

ACKNOWLEDGMENTS

We thank Alice Middleton, Marianne Goodchild, and Sandrine Touchard for assistance in a number of ways, Eirwen Morgan for fruitful discussions, and Jay Hinton, Arthur Thompson, and Saccha Luccinini at the Institute of Food Research (IFR) for assistance in establishing the microarray.

The work was funded by the European Union (RTD projects FP5-CT-2000-01126 SALARRAY and FP6-CT-2003 50523 SUPASALVAC).

REFERENCES

1. **Aldridge, P., and K. T. Hughes.** 2002. Regulation of flagellar assembly. Curr. Opin. Microbiol. **5:**160–165.
2. **Aldridge, P. D., et al.** 2006. The flagellar-specific transcription factor, sigma28, is the type III secretion chaperone for the flagellar-specific anti-sigma28 factor FlgM. Genes Dev. **20:**2315–2326.
3. **Anonymous.** 1998. PHLS evidence to House of Commons Select Committee on Agriculture; enquiry into food safety. Fourth Report for the session 1997–98. Her Majesty's Stationery Office, London, England. http://www.publications.parliament.uk/pa/cm199798/cmselect/cmagric/331iv/ag0402.htm.
4. **Anonymous.** 2006. Trends and sources of zoonoses, Zoonotic agents and antimicrobial resistance in the European Union in 2004. European Food Safety Authority, Parma, Italy. http://www.efsa.europa.eu/en/efsajournal/doc/310ar.pdf.
5. **Barrow, P. A., J. O. Hassan, and A. Berchieri, Jr.** 1990. Reduction in faecal excretion by chickens of *Salmonella typhimurium* by immunization with avirulent mutants of *S. typhimurium.* Epidemiol. Infect. **104:**413–426.
6. **Barrow, P. A., and M. A. Lovell.** 1991. Experimental infection of egg-laying hens with *Salmonella enteritidis*. Avian Pathol. **20:**339–352.
7. **Beal, R. K., C. Powers, T. F. Davison, P. A. Barrow, and A. L. Smith.** 2006. Clearance of enteric *Salmonella enterica* serovar Typhimurium in chickens is independent of B-cell function. Infect. Immun. **74:**1442–1444.
8. **Benjamini, Y., and Y. Hochberg.** 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. **57:**289–300.
9. **Bijlsma, J. J., and E. A. Groisman.** 2005. The PhoP/PhoQ system controls the intramacrophage type three secretion system of *Salmonella enterica*. Mol. Microbiol. **57:**85–96.
10. **Brinsmade, S. R., T. Paldon, and J. C. Escalante-Semerena.** 2005. Minimal functions and physiological conditions required for growth of *Salmonella enterica* on ethanolamine in the absence of the metabolosome. J. Bacteriol. **187:**8039–8046.
11. **Buchmeier, N. A., C. J. Lipps, M. Y. So, and F. Heffron.** 1993. Recombination-deficient mutants of *Salmonella typhimurium* are avirulent and sensitive to the oxidative burst of macrophages. Mol. Microbiol. **7:**933–936.
12. **Bustamante, V. H., et al.** 2008. HilD-mediated transcriptional cross-talk between SPI-1 and SPI-2. Proc. Natl. Acad. Sci. U. S. A. **105:**14591–14596.
13. Reference deleted.
14. **Cha, M. K., W. C. Kim, C. J. Lim, K. Kim, and L. H. Kim.** 2004. *Escherichia coli* periplasmic thiol peroxidase acts as lipid hydroperoxide peroxidase and the principal antioxidative function during anaerobic growth. J. Biol. Chem. **279:**8769–8778.
15. **Chang, D. E., et al.** 2004. Carbon nutrition of Escherichia coli in the mouse intestine. Proc. Natl. Acad. Sci. U. S. A. **101:**7427–7432.
16. **Clayton, D. J., et al.** 2008. Analysis of the role of 13 major fimbrial subunits in colonisation of the chicken intestine by Salmonella enteric serovar Enteritidis reveals a role for a novel locus. BMC Microbiol. **8:**228.
17. **Coates, M. E., M. E. Gregory, J. W. G. Porter, and A. P. Williams.** 1963. Vitamin B and its analogues in the gut contents of germ-free and conventional chicks. Proc. Nutr. Soc. **22:**27–35.
18. **Corbin, B. D., Y. Wang, T. K. Beuria, and W. Margolin.** 2007. Interaction between cell division proteins FtsE and FtsZ. J. Bacteriol. **189:**3026–3035.
19. **Costa, C. S., and D. N. Anton.** 1993. Round-cell mutants of *Salmonella typhimurium* produced by transposition mutagenesis: lethality of *rodA* and *mre* mutations. Mol. Gen. Genet. **236:**387–394.
20. **Craven, S. E.** 1994. Altered colonizing ability for the ceca of broiler chicks by lipopolysaccharide-deficient mutants of *Salmonella typhimurium*. Avian Dis. **38:**401–408.
21. **Dunkley, K. D., et al.** 2009. Food-borne Salmonella ecology in the avian gastro-intestinal tract. Anaerobe **15:**26–35.
22. **Edelman, S., S. Leskela, E. Ron, J. Apajalahti, and T. K. Korhonen.** 2003. *In vitro* adhesion of an avian pathogenic *Escherichia coli* O78 strain to surfaces of the chicken intestinal tract and to ileal mucus. Vet. Microbiol. **91:**41–56.
23. **Ellermeier, J. R., and J. M. Slauch.** 2007. Adaptation to the host environment: regulation of the SPI-1 type III secretion system in *Salmonella enterica* serovar Typhimurium. Curr. Opin. Microbiol. **10:**24–29.
24. **Eriksson, S., S. Lucchini, A. Thompson, M. Rhen, and J. C. Hinton.** 2003. Unravelling the biology of macrophage infection by gene expression profiling of intracellular *Salmonella enterica*. Mol. Microbiol. **47:**103–118.
25. **Fardini, Y., et al.** 2007. The YfgL lipoprotein is essential for type III secretion system expression and virulence of *Salmonella enterica* serovar Enteritidis. Infect. Immun. **75:**358–370.
26. **Foley, S. L., et al.** 2006. Comparison of subtyping methods for differentiating *Salmonella enterica* serovar Typhimurium isolates obtained from food animal sources. J. Clin. Microbiol. **44:**3569–3577.
27. **Foster, J. W.** 1999. When protons attack: microbial strategies of acid adaptation. Curr. Opin. Microbiol. **2:**170–174.
28. **Foster, J. W.** 2004. *Escherichia coli* acid resistance: tales of an amateur acidophile. Nat. Rev. Microbiol. **2:**898–907.
29. **Foster, J. W., and H. K. Hall.** 1990. Adaptive acidification tolerance response of *Salmonella typhimurium*. J. Bacteriol. **172:**771–778.
30. **Gibbs, K. A., et al.** 2004. Complex spatial distribution and dynamics of an abundant *Escherichia coli* outer membrane protein, LamB. Mol. Microbiol. **53:**1771–1783.
31. **Gophna, U., et al.** 2001. Curli fibers mediate internalization of *Escherichia coli* by eukaryotic cells. Infect. Immun. **69:**2659–2665.
32. **Hale, C. A., and P. A. de Boer.** 1999. Recruitment of ZipA to the septal ring of *Escherichia coli* is dependent on FtsZ and independent of FtsA. J. Bacteriol. **181:**167–176.
33. **Hammelman, T. A., et al.** 1996. Identification of a new *prp* locus required for propionate catabolism in *Salmonella typhimurium* LT2. FEMS Microbiol. Lett. **137:**233–239.

34. **Hansen, A. M., et al.** 2005. SspA is required for acid resistance in stationary phase by down-regulation of H-NS in *Escherichia coli*. Mol. Microbiol. **56:**719–734.
35. **Haugen, S. P., W. Ross, and R. L. Gourse.** 2008. Advances in bacterial promoter recognition and its control by factors that do not bind DNA. Nat. Rev. Microbiol. **6:**507–519.
36. **Heithoff, D. M., R. L. Sinsheimer, D. A. Low, and M. J. Mahan.** 1999. An essential role for DNA adenine methylation in bacterial virulence. Science **284:**967–970.
37. **Henry, T., et al.** 2006. The *Salmonella* effector protein PipB2 is a linker for kinesin-1. Proc. Natl. Acad. Sci. U. S. A. **103:**13497–13502.
38. **Horswill, A. R., and J. C. Escalante-Semerna.** 1997. Propionate catabolism in *Salmonella typhimurium* LT2: two divergently transcribed units comprise the *prp* locus at 8.5 centisomes, *prpR* encodes a member of the sigma-54 family of activators, and the *prpBCDE* genes constitute an operon. J. Bacteriol. **179:**928–940.
39. **Howells, A. M., et al.** 2002. Role of trehalose biosynthesis in environmental survival and virulence of *Salmonella enterica* serovar Typhimurium. Res. Microbiol. **153:**281–287.
40. **Humphries, A. D., et al.** 2003. The use of flow cytometry to detect expression of subunits encoded by 11 *Salmonella enterica* serotype Typhimurium fimbrial operons. Mol. Microbiol. **48:**1357–1376.
41. **Iwama, T., et al.** 2000. Mutational analysis of ligand recognition by Tcp, the citrate chemoreceptor of *Salmonella enterica* serovar Typhimurium. J. Bacteriol. **182:**1437–1441.
42. **Janakiraman, A., and J. M. Slauch.** 2000. The putative iron transport systems SitABCD encoded on SPI1 is required for full virulence of *Salmonella typhimurium*. Mol. Microbiol. **35:**1146–1155.
43. **Jones, M. A., S. D. Hulme, P. A. Barrow, and P. Wigley.** 2007. The Salmonella pathogenicity island 1 and Salmonella pathogenicity island 2 type III secretion systems play a major role in pathogenesis of systemic disease and gastrointestinal tract colonization of *Salmonella enterica* serovar Typhimurium in the chicken. Avian Pathol. **36:**199–203.
44. **Jones, S. A., et al.** 2007. Respiration of *Escherichia coli* in the mouse intestine. Infect. Immun. **75:**4891–4899.
45. **Jones, S. A., et al.** 2008. Glycogen and maltose utilization by *Escherichia coli* O157:H7 in the mouse intestine. Infect. Immun. **76:**2531–2540.
46. **Kaiser, P., et al.** 2000. Differential cytokine expression in avian cells in response to invasion by *Salmonella typhimurium*, *Salmonella enteritidis* and *Salmonella gallinarum*. Microbiology **146:**3217–3226.
47. **Keener, J., and M. Nomura.** 1996. Regulation of ribosome synthesis, p. 1417–1431. *In* F. C. Neidhardt, et al. (ed.) *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, DC.
48. **Kieboom, J., and T. Abee.** 2006. Arginine-dependent acid resistance in *Salmonella enterica* serovar Typhimurium. J. Bacteriol. **188:**5650–5653.
49. **Klumpp, S., Z. Zhang, and T. Hwa.** 2009. Growth-rate dependent global effects on gene expression in bacteria. Cell **139:**1366–1375.
50. **Knodler, L. A., et al.** 2002. Salmonella effectors within a single pathogenicity island are differentially expressed and translocated by separate type III secretion systems. Mol. Microbiol. **43:**1089–1103.
51. **Kruse, T., J. Moller-Jensen, A. Lobner-Olesen, and K. Gerdes.** 2003. Dysfunctional MreB inhibits chromosome segregation in *Escherichia coli*. EMBO J. **22:**5283–5292.
52. **Latasa, C., et al.** 2005. BapA, a large secreted protein required for biofilm formation and host colonization of *Salmonella enterica* serovar Enteritidis. Mol. Microbiol. **58:**1322–1339.
53. **Lewis, N. E., et al.** 2010. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol. Syst. Biol. **6:**390.
54. **Lucas, R. L., and C. A. Lee.** 2001. Roles of *hilC* and *hilD* in regulation of *hilA* expression in *Salmonella enterica* serovar Typhimurium. J. Bacteriol. **183:**2733–2745.
55. Reference deleted.
56. **McClelland, M., et al.** 2001. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. Nature **413:**852–856.
57. **McMeechan, A., et al.** 2005. Glycogen production by different Salmonella enterica serotypes: contribution of functional glgC to virulence, intestinal colonization and environmental survival. Microbiology **151:**3969–3977.
58. **Moncrief, M. B. C., and M. E. Maguire.** 1998. Magnesium and the role of *mgtC* in growth of *Salmonella typhimurium*. Infect. Immun. **66:**3802–3809.
59. **Morgan, E., et al.** 2004. Identification of host-specific colonization factors of *Salmonella enterica* serovar Typhimurium. Mol. Microbiol. **54:**994–1010.
60. **Palacios, S., and J. C. Escalante-Semerena.** 2000. *prpR*, *ntrA*, and *ihf* functions are required for expression of the *prpBCDE* operon, encoding enzymes that catabolize propionate in *Salmonella enterica* serovar Typhimurium LT2. J. Bacteriol. **182:**905–910.
61. **Palacios, S., V. J. Starai, and J. C. Escalante-Semerena.** 2003. Propionyl coenzyme A is a common intermediate in the 1,2-propanediol and propionate catabolic pathways needed for expression of the *prpBCDE* operon during growth of *Salmonella enterica* on 1,2-propanediol. J. Bacteriol. **185:**2802–2810.
62. **Pichoff, S., and J. Lutkenhaus.** 2007. Identification of a region of FtsA required for interaction with FtsZ. Mol. Microbiol. **64:**1129–1138.
63. **Poulsen, L. K., T. R. Licht, C. Rang, K. A. Krogfelt, and S. Molin.** 1995. Physiological state of *Escherichia coli* BJ4 growing in the large intestines of streptomycin-treated mice. J. Bacteriol. **177:**5840–5845.
64. **Price-Carter, M., J. Tingey, T. A. Bobik, and J. R. Roth.** 2001. The alternative electron acceptor tetrathionate supports B_{12}-dependent anaerobic growth of *Salmonella enterica* serovar Typhimurium on ethanolamine or 1,2-propanediol. J. Bacteriol. **183:**2463–2475.
65. **Rodrigue, D. C., R. V. Tauxe, and B. Rowe.** 1990. International increase in Salmonella enteritidis: a new pandemic? Epidemiol. Infect. **105:**21–27.
66. **Rondon, M. R., R. Kazmierczak, and J. C. Escalante-Semerena.** 1995. Glutathione is required for maximal transcription of the cobalamin biosynthetic and 1,2-propanediol utilization (*cob*/*pdu*) regulon and for the catabolism of ethanolamine, 1,2-propanediol, and propionate in *Salmonella typhimurium* LT2. J. Bacteriol. **177:**5434–5439.
67. **Roof, D. M., and J. R. Roth.** 1992. Autogenous regulation of ethanolamine utilization by a transcriptional activator of the *eut* operon in *Salmonella typhimurium*. J. Bacteriol. **174:**6634–6643.
68. **Sheppard, D. E., J. T. Penrod, T. Bobik, E. Kofoid, and J. R. Roth.** 2004. Evidence that a B_{12}-adenosyl transferase is encoded within the ethanolamine operon of *Salmonella enterica*. J. Bacteriol. **186:**7635–7644.
69. **Shi, Y., M. J. Cromie, F. F. Hsu, J. Turk, and E. A. Groisman.** 2004. PhoP-regulated *Salmonella* resistance to the antimicrobial peptides magainin 2 and polymyxin B. Mol. Microbiol. **53:**229–241.
70. **Smith, H. W., and J. F. Tucker.** 1980. The virulence of salmonella strains for chickens: their excretion by infected chickens. J. Hyg. (Lond.) **84:**479–488.
71. **Smyth, G. K.** 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. **3:**article 3.
72. **Smyth, G. K., J. Michaud, and H. Scott.** 2005. The use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics **21:**2067–2075.
73. **Snoeyenbos, G. H.** 1991. Pullorum disease, p. 73–87. *In* B. W. Calnek (ed.), Diseases of poultry, 9th ed. Iowa State University Press, Ames, IA.
74. **Stecher, B., M. Barthel, M. C. Schlumberger, M. Kremer, and W.-D. Hardt.** 2008. Motility allows S. Typhimurium to benefit from the mucosal defence. Cell. Microbiol. **10:**1166–1180.
75. **Stojiljkovic, I., A. J. Baumler, and F. Heffron.** 1995. Ethanolamine utilization in *Salmonella typhimurium*: nucleotide sequence, protein expression, and mutational analysis of the *cchA cchB eutE eutJ eutH* gene cluster. J. Bacteriol. **177:**1357–1366.
76. **Strom, A. R., and I. Kaasen.** 1993. Trehalose metabolism in *Escherichia coli*: stress protection and stress regulation of gene expression. Mol. Microbiol. **8:**205–210.
77. **Tamai, E., T. Shimamoto, M. Tsuda, T. Mizushima, and T. Tsuchiya.** 1998. Conversion of temperature-sensitive to -resistant gene expression due to mutations in the promoter region of the melibiose operon in *Escherichia coli*. J. Biol. Chem. **273:**16860–16864.
78. **Tinker, J. K., L. S. Hancox, and S. Clegg.** 2001. FimW is a negative regulator affecting type 1 fimbrial expression in *Salmonella enterica* serovar Typhimurium. J. Bacteriol. **183:**435–442.
79. **Tinker, J. K., and S. Clegg.** 2000. Characterization of FimY as a coactivator of type 1 fimbrial expression in *Salmonella enterica* serovar Typhimurium. Infect. Immun. **68:**3305–3313.
80. Reference deleted.
81. **Tsang, A. W., A. R. Horswill, and J. C. Escalante-Semerena.** 1998. Studies of regulation of expression of the propionate (*prpBCDE*) operon provide insights into how *Salmonella typhimurium* LT2 integrates its 1,2-propanediol and propionate catabolic pathways. J. Bacteriol. **180:**6511–6518.
82. **Turner, A. K., M. A. Lovell, S. D. Hulme, L. Zhang-Barber, and P. A. Barrow.** 1998. Identification of *Salmonella typhimurium* genes required for colonization of the chicken alimentary tract and for virulence in newly hatched chicks. Infect. Immun. **66:**2099–2106.
83. **Turner, A. K., et al.** 2003. Contribution of proton-translocating proteins to the virulence of *Salmonella enterica* serovars Typhimurium, Gallinarum, and Dublin in chickens and mice. Infect. Immun. **71:**3392–3401.
84. **Tweeddale, H., L. Notley-McRobb, and T. Ferenci.** 1998. Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. J. Bacteriol. **180:**5109–5116.
85. **Van Bibber, M., C. Bradbeer, N. Clark, and J. R. Roth.** 1999. A new class of cobalamin transport mutants (*btuF*) provides genetic evidence for a periplasmic binding protein in *Salmonella typhimurium*. J. Bacteriol. **181:**5539–5541.
86. **Varma, A., M. dePedro, and K. D. Young.** 2007. FtsZ directs a second mode of peptidoglycan synthesis in *Escherichia coli*. J. Bacteriol. **189:**5692–5704.
87. **Watson, M.** 2005. ProGenExpress: visualization of quantitative data on prokaryotic genomes. BMC Bioinformatics **6:**98.
88. **Weart, R. B., and P. A. Levin.** 2003. Growth rate-dependent regulation of medial FtsZ ring formation. J. Bacteriol. **185:**2826–2834.
89. **Wigley, P., M. A. Jones, and P. A. Barrow.** 2002. *Salmonella enterica* serovar

Pullorum requires the *Salmonella* pathogenicity island 2 type III secretion system for virulence and carriage in chickens. Avian Pathol. **31:**501–506.
90. **Winter, S. E., et al.** 2010. Gut inflammation provides a respiratory electron acceptor for Salmonella. Nature **467:**426–429.
91. **Withanage, G. S. K., et al.** 2004. Rapid expression of chemokines and proinflammatory cytokines in newly hatched chickens infected with *Salmonella enterica* serovar Typhimurium. Infect. Immun. **72:**2152–2159.
92. **Woodall, C. A., et al.** 2005. *Campylobacter jejuni* gene expression in the chick cecum: evidence for adaptation to a low-oxygen environment. Infect. Immun. **73:**5278–5285.
93. **Wu, W.-F., M. L. Urbanowski, and G. V. Stauffer.** 1992. Role of the MetR regulatory system in vitamin B_{12}-mediated repression of the *Salmonella typhimurium metE* gene. J. Bacteriol. **174:**4833–4837.
94. **Wu, W.-F., M. L. Urbanowski, and G. V. Stauffer.** 1995. Characterization of a second MetR-binding site in the *metE metR* regulatory region of *Salmonella typhimurium*. J. Bacteriol. **177:**1834–1839.
95. **Yamamoto, S., and K. Kutsukake.** 2006. FljA-mediated posttranscriptional control of phase 1 flagellin expression in flagellar phase variation of *Salmonella enterica* serovar Typhimurium. J. Bacteriol. **188:**958–967.
96. **Yang, Y. H., et al.** 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. **30:**e15.
97. **Yohannes, E., A. E. Thurber, J. C. Wilks, D. P. Tate, and J. L. Slonczewski.** 2005. Polyamine stress at high pH in *Escherichia coli* K-12. BMC Microbiol. **5:**59–66.
98. **Zaharik, M. L., et al.** 2004. The *Salmonella enterica* serovar Typhimurium divalent cation transport systems MntH and SitABCD are essential for virulence in an *Nramp1*^{G169} murine typhoid model. Infect. Immun. **72:**5522–5525.
99. Reference deleted.
100. **Zhang-Barber, L., et al.** 1997. Influence of genes encoding proton-translocating enzymes on suppression of *Salmonella typhimurium* growth and colonization. J. Bacteriol. **179:**7186–7190.
101. **Zhou, D., W.-D. Hardt, and J. E. Galan.** 1999. Salmonella typhimurium encodes a putative iron transport system within the centisome 63 pathogenicity island. Infect. Immun. **67:**1974–1981.

Editor: A. J. Bäumler

BMC Bioinformatics

Software

ProGenExpress: Visualization of quantitative data on prokaryotic genomes

Michael Watson*

Address: Institute for Animal Health, Compton laboratory, High street, Compton, Newbury, RG20 7NN, UK

Email: Michael Watson* - michael.watson@bbsrc.ac.uk

* Corresponding author

Published: 13 April 2005

BMC Bioinformatics 2005, **6**:98 doi:10.1186/1471-2105-6-98

Received: 09 February 2005

Accepted: 13 April 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/98

Abstract

Background: The integration of genomic information with quantitative experimental data is a key component of systems biology. An increasing number of microbial genomes are being sequenced, leading to an increasing amount of data from post-genomics technologies. The genomes of prokaryotes contain many structures of interest, such as operons, pathogenicity islands and prophage sequences, whose behaviour is of interest during infection and disease. There is a need for simple and novel tools to display and analyse data from these integrated datasets, and we have developed ProGenExpress as a tool for visualising arbitrarily complex numerical data in the context of prokaryotic genomes.

Results: Here we describe ProGenExpress, an R package that allows researchers to easily and quickly visualize quantitative measurements, such as those produced by microarray experiments, in the context of the genome organization of sequenced prokaryotes. Data from microarrays, proteomics or other whole-genome technologies can be accurately displayed on the genome. ProGenExpress can also search for novel regions of interest that consist of groups of adjacent genes that show similar patterns across the experimental data set. We demonstrate ProGenExpress with microarray data from a time-course experiment involving *Salmonella typhimurium*.

Conclusion: ProGenExpress can be used to visualize quantitative data from complex experiments in the context of the genome of sequenced prokaryotes, and to find novel regions of interest.

Background

The genomes of prokaryotic organisms contain many structures that may be involved in pathogenicity, including a variety of operons, pathogenicity islands and prophage sequences. Operons are sets of adjacent genes in bacteria that form a single transcriptional unit, and many, such as those coding for flagella [1] or fimbriae [2], have been implicated in pathogenicity. Pathogenicity islands are distinct regions of the genome that confer virulence upon the host, and are found in many pathogens of humans, animals and plants, and at least ten pathogenicity islands have been identified in *Salmonella* alone [3]. Prophage sequences represent the chromosomes of bacteriophage integrated as part of the genome of the bacterial host, and have also been implicated in pathogenicity in several species [4].

In order to study the behaviour of these elements, it is essential to integrate information about the genome structure of an organism with quantitative measurements

produced by post-genomic technologies, such as those from microarray or proteomics experiments. This integrative biology approach is a key feature of systems biology. Studying the behaviour of these genomic elements, and other groups of adjacent genes, during infection and disease may reveal important information about the molecular mechanisms underlying pathogenicity.

Several microbial genome viewers have been developed which allow quantitative data to be displayed on the genome. The **Microbial Genomes Viewer** [5] offers a good online solution, however users must install a browser plug-in and may not be comfortable transmitting data over the internet. **GenoMap** [6] can be used to create plots of microarray data on microbial genomes, and is available as Tcl/TK source code. **Genome2D** [7] also offers good visualisation of quantitative data on microbial genomes, but is limited to the Windows operating system. Finally, **GenomeViz** [8] has recently been released, which offers much functionality, including visualisation of quantitative data, genome alignments and GC content. However this software is currently limited to unix-based systems. **All** of the above solutions are limited in two respects. Firstly, the quantitative values are represented as a colour-scale, which reduces the accuracy of the data and which may present problems in comparing one colour to the next. Secondly, the above tools can only display a single value for each gene, which precludes the visualisation of more complex data, such as a time-course experiment.

Implementation

ProGenExpress is released as a package for R. R is a freely available, open-source statistical package [9] that is widely used in the biological community. R has very powerful statistical and graphical capabilities, and many add-on packages are freely available. The bioconductor project [10,11] provides a huge number of add-on packages for R, covering a wide range of biological data analysis applications, and the implementation of ProGenExpress in R provides seamless integration with many of these packages. ProGenExpress is written in the native R language and has been fully tested on both windows and linux. R is available for windows, linux, unix and MacOS (including MacOS X).

Results and discussion

ProGenExpress has been written to allow researchers to quickly and simply visualize the behaviour of bacterial genomic regions of any size during experiments using whole genome technologies, such as microarray or proteomics experiments. For information relating to the genome organisation of prokaryotes, ProGenExpress includes functions for downloading and reading both NCBI .ptt files, which describe the location of protein coding genes in bacteria in a tabular format, and include links to the COGs database [12], and whole genome RefSeq entries [13]. For the quantitative experimental data, ProGenExpress can use the objects created by many of the packages from the bioconductor project [10,11], or data imported into R from text files, SQL databases and Excel.

There are currently 225 completed prokaryotic genomes in RefSeq [15] that ProGenExpress can read, and though the utility of ProGenExpress is demonstrated here using microarray data, any kind of numerical data that can be linked to the genes of prokaryotic organisms can be displayed using ProGenExpress. Where measures of the statistical significance of the data points for each gene are available, these can be passed to the plotting functions of ProGenExpress, with the result that those genes that are not significant will be plotted in white and those that are significant will be plotted in their normal plotting colour.

The genome is represented as two barplots, one for each strand. Each gene has a number of bars equal to the number of experimental data sets included, allowing time-course or complex strain/treatment experiments to be plotted. Distance between the bars for each gene is representative of intergenic distance. Slices of the genome can be selected either by base range, gene synonym or gene name. Both horizontal and vertical plots are possible, and bars can be coloured either by numerical value or by COGs [12] functional category.

The software is demonstrated here using microarray data from Eriksson *et al* [14]. This data set consists of gene expression measurements from intracellular *Salmonella typhimurium* at 4, 8 and 12 hours post murine macrophage infection. Gene expression values were calculated as the relative expression level of test RNA to that of RNA from bacteria grown *in vitro*, and the data has been centred and normalised according to Eriksson *et al* [14]. Data from Erikson *et al* is available as a spreadsheet [14]. This spreadsheet was pre-processed to contain only columns for gene synonym, gene name and relative expression level of test RNA to control RNA on a log 2 scale for each of the three time points. The spreadsheet was saved as tab-delimited text and read in to R using the read.table() function. The S typhimurium genome and plasmid sequences were read in to R using the read.ptt() function, with RefSeq files NC_003197.ptt and NC_003277.ptt respectively. The microarray data was linked to the gene location data using the linkem.avg() function. Images of the microarray data on both the entire genome and the plasmid were then generated using the plotrange() and plotrange.vertical() functions in conjunction with jpeg(), an internal R function. The results were viewed in Internet Explorer. Finally, the find.region() function was used to find regions of interest as described below.

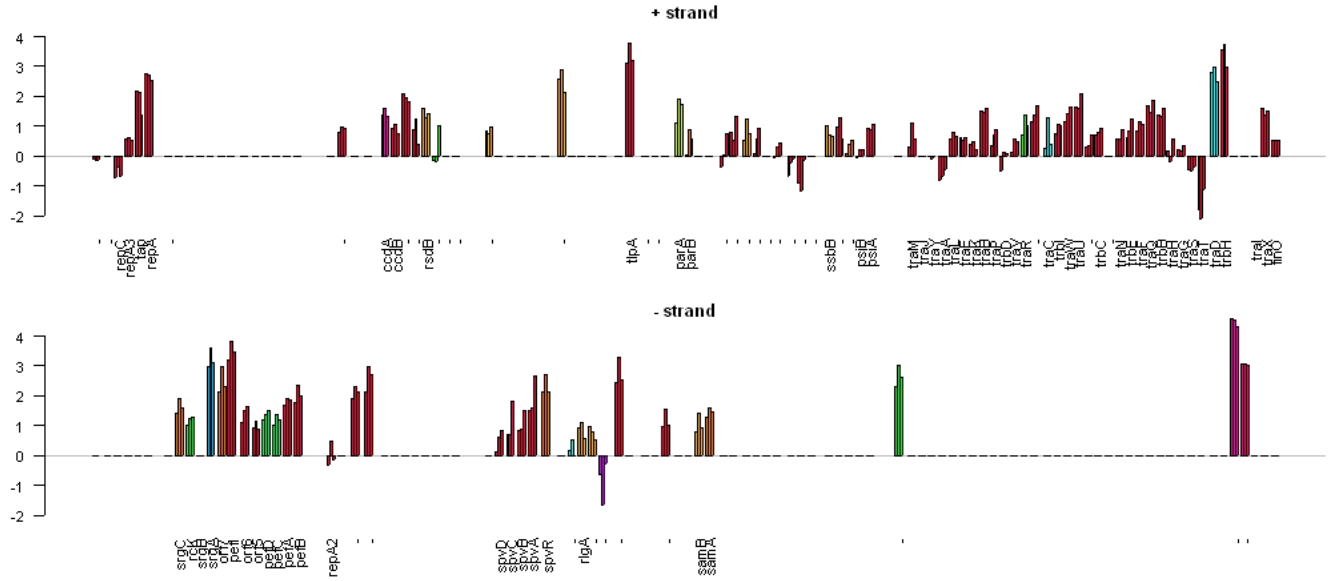


Figure 1

Gene expression measurements for *Salmonella typhimurium* plasmid pSLT. Gene expression measurements for the pSLT plasmid of *Salmonella typhimurium*. The genome is represented as two barplots, one for each strand. Each gene has three bars representing expression at 4 h, 8 h and 12 h post macrophage infection. Gene expression measurements (y-axis) are ratios of test RNA to control RNA on the log 2 scale. Bars are colour-coded according to the COGs [12] functional category. Distance between genes is relative and representative of intergenic distance. This image clearly shows that the majority of genes on the plasmid are up-regulated during macrophage infection.

Figure 1 shows the expression of all genes on *Salmonella typhimurium* LT2 plasmid pSLT, coloured by COGs functional category. The majority of the genes on this plasmid are up-regulated at all three time-points, implying a role for this plasmid during macrophage infection. Figure 2 displays a smaller region of the genome containing the fli operon, with all genes in the operon displaying similar expression profiles. Erikksson *et al* [14] found 919 genes to be significantly differentially expressed, and that measure of statistical significance has been incorporated into Figure 2. Significant genes are coloured normally, whereas those that are not significant are white. All but three of the 14 genes in the operon are shown to be significantly differentially expressed, suggesting that the whole operon is differentially expressed and that perhaps the measure of statistical significance used is too stringent. Finally, Figure 3 is a vertical plot of *Salmonella* pathogenicity island II (SPI-II), showing that most genes on this island are up-regulated at all three time-points. This island encodes a type III secretion system, and has been shown to be required for systemic infection by facilitating replication of intracellular bacteria within membrane-bound Salmonella-containing vacuoles [3].

ProGenExpress can also search for operons and other regions of interest by looking for clusters of genes that are close together and which display similar patterns in the experimental data. Using this facility, we identified over 200 potential regions of interest in *Salmonella typhimurium* consisting of four genes or more, including several known operons and potential unannotated operons. Figure 4 shows a region of the genome containing a group of six genes that has been found using ProGenExpress. The genes have no assigned gene name, have either an unknown or putative/predicted function, are close together on the genome and have similar expression profiles across the three time-points. We believe these genes may represent an unannotated operon.

ProGenExpress has several advantages over existing software. The package seamlessly integrates with the bioconductor project and the many packages available in R for microarray analysis, including limma, marray and affy, and is available for both Windows and Linux, amongst others. Both horizontal and vertical plots are possible, and an unlimited number of data points for each gene can be plotted, allowing for the visualization and analysis of

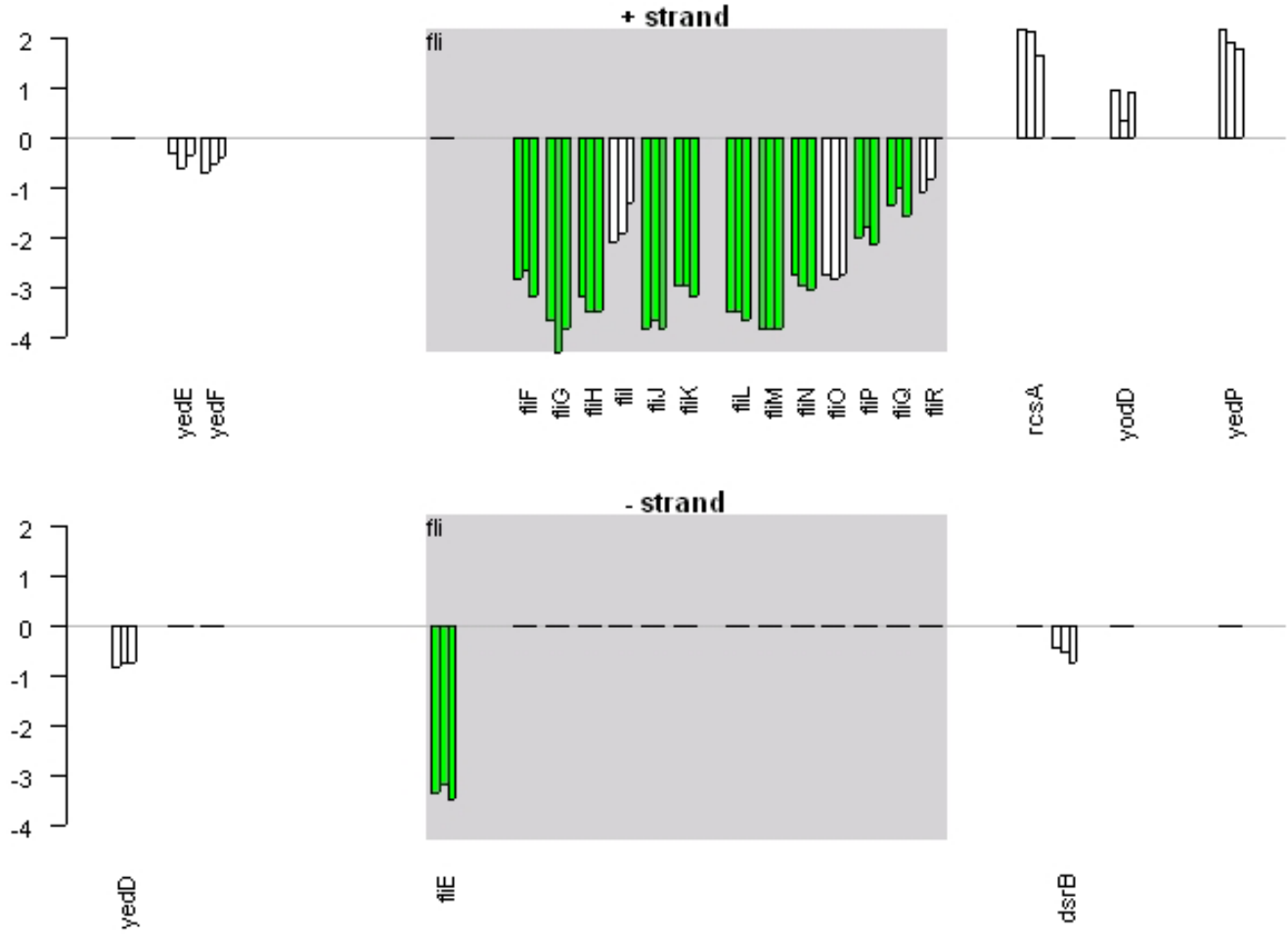


Figure 2

Gene expression measurements for flagella biosynthesis. Gene expression measurements for the fli operon from *Salmonella typhimurium*. The genome is represented as two barplots, one for each strand. Each gene has three bars representing expression at 4 h, 8 h and 12 h post macrophage infection. Gene expression measurements (y-axis) are ratios of test RNA to control RNA on the log 2 scale. Distance between genes is relative and representative of intergenic distance. The fli operon is involved in flagella biosynthesis and this image clearly shows that the entire operon is strongly down-regulated during macrophage infection. Significantly differentially expressed genes are coloured red or green, whereas non-significant genes are coloured white. All but three of the 14 genes in the operon are significantly differentially expressed.

complex time course or strain/treatment experiments. Furthermore, the bar-plots display numerical data accurately, and do not rely on a colour-scale to depict values. Finally, the ability to search integrated genomic and post-genomic data sets for clusters of genes which behave similarly represents an opportunity for the discovery of novel genomic elements involved in pathogenicity.

Conclusion

We describe ProGenExpress, an open-source R package which allows researchers to quickly and easily visualise quantitative data from arbitrarily complex experiments in the context of the genome of sequenced prokaryotes. ProGenExpress can also be used to search for genomic regions which may represent coherent functional units. We show how ProGenExpress can be used to visualise microarray data from a time-course experiment on the genome of *Salmonella typhimurium*, and to find unannotated genomic regions that may be involved in pathogenicity. Future plans for the software include the ability to read data from ensembl databases, and the development of visualisation

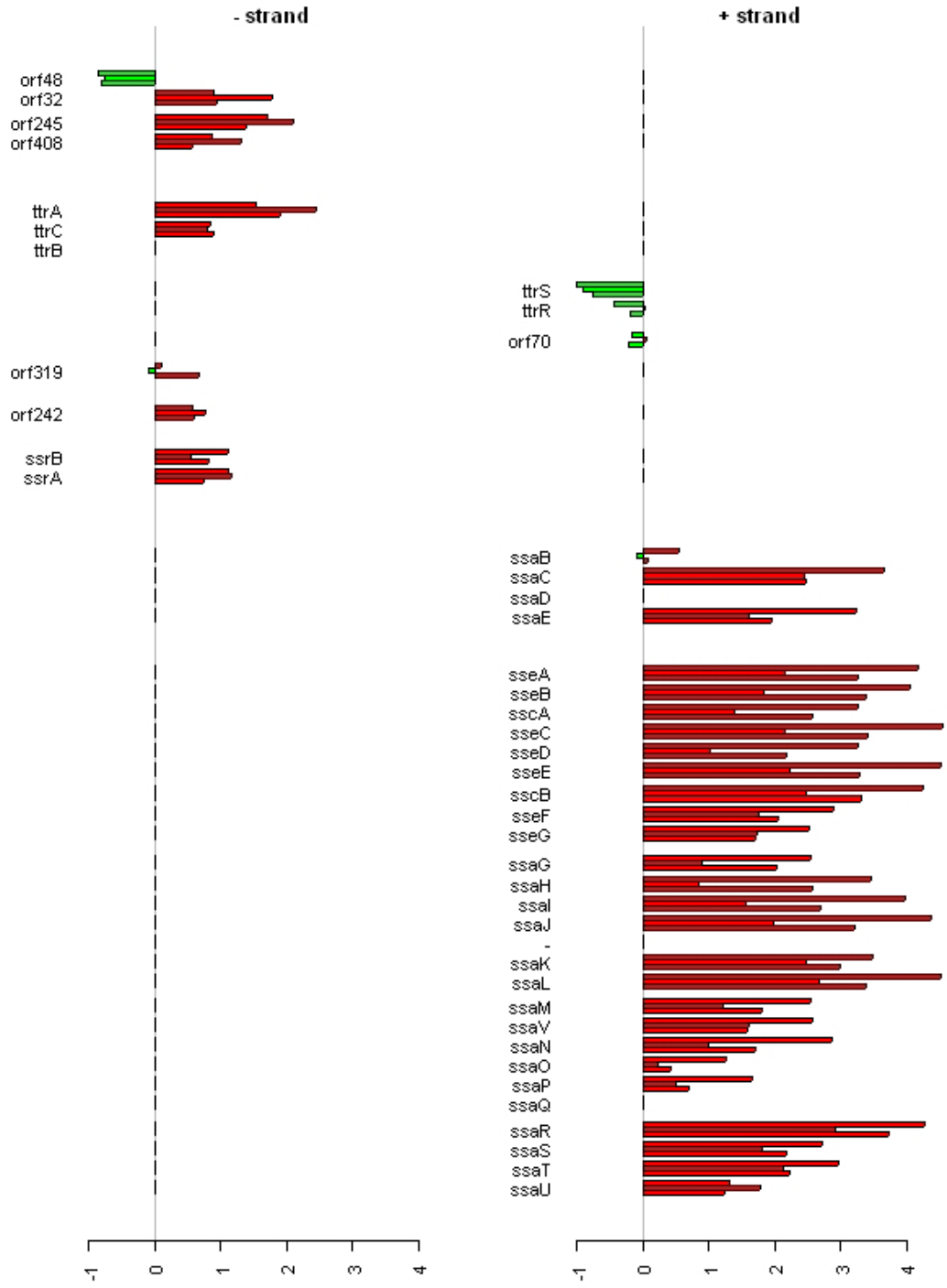


Figure 3

Gene expression measurements for Salmonella pathogenicity island II. Gene expression measurements for Salmonella Pathogenicity Island II (SPI-II) from *Salmonella typhimurium*. The genome is represented as two barplots, one for each strand. Each gene has three bars representing expression at 4 h, 8 h and 12 h post macrophage infection. Gene expression measurements (y-axis) are ratios of test RNA to control RNA on the log 2 scale. Distance between genes is relative and representative of intergenic distance. This island has been linked to pathogenicity, and encodes a type III secretion system. It is required for systemic infection and intracellular pathogenesis by facilitating replication of intracellular bacteria within membrane-bound Salmonella-containing vacuoles [3]. Here we can clearly see that the majority of genes in the island are strongly up-regulated during macrophage infection.

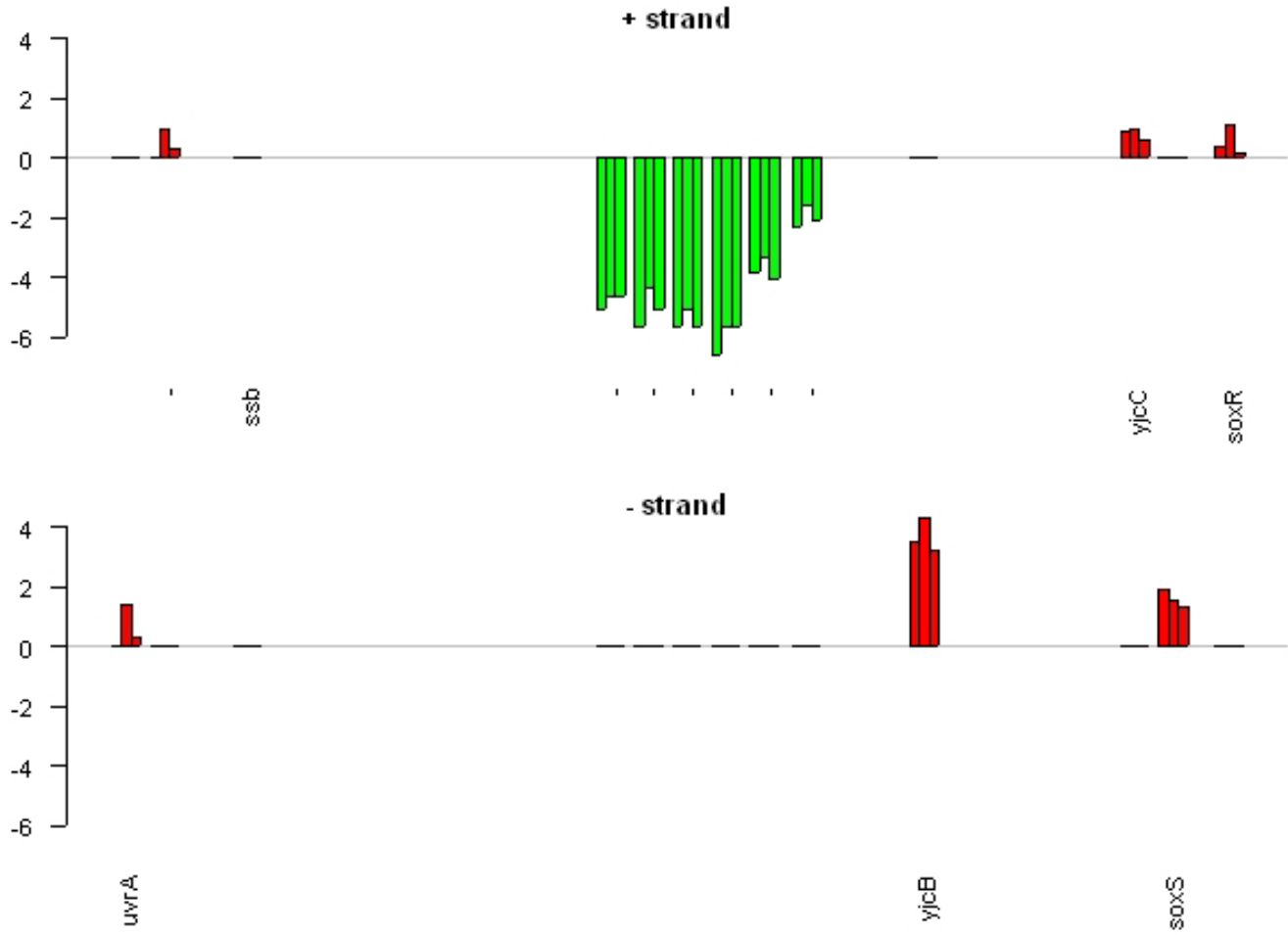


Figure 4
A putative operon. Gene expression measurements from a genomic region of *Salmonella typhimurium*. The genome is represented as two barplots, one for each strand. Each gene has three bars representing expression at 4 h, 8 h and 12 h post macrophage infection. Gene expression measurements (y-axis) are ratios of test RNA to control RNA on the log 2 scale. Distance between genes is relative and representative of intergenic distance. The image shows six genes that have been identified by ProGenExpress as potentially interesting. The genes are very close together on the genome and display similar expression patterns, and therefore could represent an as yet unannotated operon. The genes have synonyms STM4257 – STM4262 and currently have no confirmed function.

tools for eukaryotic genomes. Software updates and new releases will be available from the project home page.

Availability and requirements

- **Project Name:** ProGenExpress
- **Project Home Page**: http://progenexpress.sf.net
- **Operating Systems:** Windows, Linux, Unix
- **Programming Language:** R
- **Other Requirements:** R version 2.0 or above
- **License:** GNU GPL

Authors' contributions

MW developed and tested the software in full.

List of abbreviations

COG: Cluster of Orthologous Groups

SPI-II: *Salmonella* pathogenicity island II

Acknowledgements

This work was funded by the core strategic grant of the Institute for Animal Health, provided by the BBSRC.

References

1. Robertson JMC, McKenzie NH, Duncan M, Allen-Vercoe E, Woodward MJ, Flint HJ, Grant G: **Lack of flagella disadvantages Salmonella enterica serovar Enteritidis during the early stages of infection in the rat.** *J Med Micro* 2003, **52:**91-99.
2. Baumler AJ, Tsolis RM, Heffron F: **Contribution of fimbrial operons to attachment to and invasion of epithelial cell lines by Salmonella typhimurium.** *Infect Immun* 1996, **64:**1862-65.
3. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, O'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG: **Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18.** *Nature* 2001, **413(6858):**848-52.
4. Banks DJ, Lei B, Musser JM: **Prophage induction and expression of prophage-encoded virulence factors in group A Streptococcus serotype M3 strain MGAS315.** *Infect Immun* 2003, **71(12):**7079-86.
5. Kerkhoven R, van Enckevort FH, Boekhorst J, Molenaar D, Siezen RJ: **Visualization for genomics: the Microbial Genome Viewer.** *Bioinformatics* 2004, **20(11):**1812-14.
6. Sato N, Ehira S: **GenoMap, a circular genome data viewer.** *Bioinformatics* 2003, **19(12):**1583-84.
7. Baerends RJ, Smits WK, de Jong A, Hamoen LW, Kok J, Kuipers OP: **Genome2D: a visualization tool for the rapid analysis of bacterial transcriptome data.** *Genome Biol* 2004, **5(5):**R37.
8. Ghai R, Hain T, Chakraborty T: **GenomeViz: visualizing microbial genomes.** *BMC Bioinformatics* **5(1):**198.
9. **R** [http://www.r-project.org]
10. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10):**R80.
11. **Bioconductor** [http://www.bioconductor.org]
12. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4(1):**41.
13. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33:**D501-4.
14. Eriksson S, Lucchini S, Thompson A, Rhen M, Hinton JCD: **Unravelling the biology of macrophage infection by gene expression profiling of intracellular Salmonella enterica.** *Molecular Microbiology* 2003, **47(1):**103-118.
15. **NCBI Refseq completed microbial genomes** [http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html]

frontiers in
GENETICS

METHODS ARTICLE
published: 05 September 2013
doi: 10.3389/fgene.2013.00168

Meta4: a web application for sharing and annotating metagenomic gene predictions using web services

*Emily J. Richardson[1], Franck Escalettes[2], Ian Fotheringham[2], Robert J. Wallace[3] and Mick Watson[1,4]**

[1] ARK-Genomics, The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Midlothian, UK
[2] Ingenza Ltd., Roslin BioCentre, Midlothian, UK
[3] Rowett Institute of Nutrition and Health, University of Aberdeen, Aberdeen, UK
[4] Edinburgh Genomics, University of Edinburgh, Edinburgh, UK

Edited by:
John Hancock, University of Cambridge, UK

Reviewed by:
Ugur Sezerman, Sabanci University, Turkey
Pascale Gaudet, Swiss Institute of Bioinformatics, Switzerland

***Correspondence:**
Mick Watson, ARK-Genomics, The Roslin Institute and R(D)SVS, University of Edinburgh, Division of Genetics and Genomics, Easter Bush, Midlothian EH25 9RG, UK
e-mail: mick.watson@roslin.ed.ac.uk

Whole-genome shotgun metagenomics experiments produce DNA sequence data from entire ecosystems, and provide a huge amount of novel information. Gene discovery projects require up-to-date information about sequence homology and domain structure for millions of predicted proteins to be presented in a simple, easy-to-use system. There is a lack of simple, open, flexible tools that allow the rapid sharing of metagenomics datasets with collaborators in a format they can easily interrogate. We present Meta4, a flexible and extensible web application that can be used to share and annotate metagenomic gene predictions. Proteins and predicted domains are stored in a simple relational database, with a dynamic front-end which displays the results in an internet browser. Web services are used to provide up-to-date information about the proteins from homology searches against public databases. Information about Meta4 can be found on the project website[1], code is available on Github[2], a cloud image is available, and an example implementation can be seen at http://www.ark-genomics.org/tools/meta4

Keywords: metagenomics, database, web service, gene discovery, bioinformatics

INTRODUCTION

Whole-genome shotgun (WGS) metagenomics can be defined as the application of high-throughput sequencing technologies to whole environmental samples, enabling scientists to assay the genomes of all organisms within a particular ecosystem, be it the human gut microbiome (Yatsunenko et al., 2012), permafrost (Mackelprang et al., 2011), or the Sargasso Sea (Venter et al., 2004). One of the aims of such endeavors is to discover novel enzymes that may have be of use to the biotechnology industry (Cowan et al., 2005), and metagenomics has been identified as a major mechanism for increasing the "sequencing space" from which to discover new biocatalysts (Cowan et al., 2004).

Whole-genome shotgun metagenomics experiments routinely produce hundreds of gigabases of sequencing data. A generalized analysis pipeline for such data is to (i) assemble the genomic data *de novo*; (ii) predict genes and proteins on the resulting contigs and scaffolds; (iii) assign domains and function to those proteins; (iv) interpret those findings within the biological context. It is not unusual for such studies to generate several million novel genes/proteins – Venter et al. (2004) reported over 1.2 million novel genes, and Hess et al. (2011) reported over 2.5 million putative genes, 27755 containing a domain of interest: those relevant to biomass degradation.

Metagenomic assembly poses specific problems over and above those of single genome assembly. The attempt to simultaneously assemble thousands of different genomes often results in large and complex assembly graphs. These require more memory to create and query, and also often require extra information in order to find true paths through the graphs. Ray Meta (Boisvert et al., 2012) is a massively distributed metagenome assembler that uses message passing, whereas Pell et al. (2012) reduce memory requirements using a bloom filter and use kmer connectivity to improve the assembly process. Other tools attempt to partition the assembly graph – Meta IDBA using graph connectivity (Peng et al., 2011) and MetaVelvet using both coverage and connectivity (Namiki et al., 2012). Finally, MetAMOS (Treangen et al., 2013) is a metagenomics pipeline that combines a number of published tools for metagenomic analysis.

Once the raw metagenomic reads have been assembled into contigs and scaffolds, the next stage is an attempt to predict the location of genes. Here again, metagenomics poses particular problems when compared to single bacterial genome annotation (recently reviewed in Richardson and Watson, 2013). Specifically, traditional bacterial gene predictors use models trained on a single, related genome; as with metagenomics we sequence thousands of genomes simultaneously, this is no longer appropriate. A number of tools have been published for metagenomic gene prediction, including MetaGeneAnnotator (Noguchi et al., 2008), Orphelia (Hoff et al., 2009), FragGeneScan (Rho et al., 2010), and Glimmer-MG (Kelley et al., 2012). Yok and Rosen (2011) propose a combination of tools.

Once genes have been annotated, domains can be assigned to protein-coding genes using traditional approaches, such as HMMER (Eddy, 2009) searches of domain databases such as Pfam (Punta et al., 2012), and the use of tools such as InterProScan (Mulder and Apweiler, 2007).

[1] http://www.ark-genomics.org/bioinformatics/meta4
[2] https://github.com/mw55309/meta4

After raw reads from metagenomics experiments have been assembled and annotated, researchers are left with a very large and rich dataset which can be difficult to query and share. Tools that allow multiple users to browse and query such datasets, either privately within a consortium, or as part of a public collaboration, remain under-developed. It is essential that simple, open, and flexible tools are provided to allow scientists to easily access the outputs of metagenomic gene discovery projects. Here we describe Meta4, a web application that is easy to install, that should work on any standard LAMP (Linux, Apache, MySQL, PHP) server, and which allows users to search and browse large collections of metagenomic gene predictions in a user-friendly web interface. In addition, Meta4 makes use of web services to provide up-to-date annotation.

There are a few existing tools for organizing and analyzing metagenomic data on the web; however, despite being feature-rich, many are closed systems. The integrated microbial genomes and metagenomes (IMG/M) system (Markowitz et al., 2012) allows comprehensive analysis of genomes and metagenomes sequenced at the Joint Genome Institute (JGI). However, the system is not open-source, it is not possible to download the code and create a local installation, the software is only extensible by the authors and it is not easy to integrate your own data – one must e-mail the authors and request integration. Similarly, the Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA; Sun et al., 2011) is a workflow-based, feature-rich website for metagenomic analysis; however, the same issues remain in that it is not open-source, it is only extensible by the authors, it is not possible to create a local installation, and users must e-mail the authors to request integration of their data. Luckily, the metagenomics RAST server (MG-RAST; Meyer et al., 2008), a very popular and comprehensive tool for metagenomic data analysis, is far more open, with users encouraged to submit their own data, and the code is available on github[3]. However, even the authors admit, local installations of the tool are difficult, they advise against it, and no support for such an undertaking is available[4].

All three tools are feature- and function-rich, and aim to be complete systems for the assembly, annotation, and comparison of multiple metagenomic samples. One problem with systems such as IMG/M and CAMERA is an inability for users to maintain data privacy; once data is uploaded to these systems, it is available for the public to see. MG-RAST does have the option to submit to a private queue, but this is a low priority queue. As such, these tools are not designed for the simple task of sharing large amounts of data quickly and simply. Meta4 is not designed to compete with these tools in terms of functionality; rather, it is a simple tool allowing the rapid sharing of metagenomic results that is easily extensible by the addition of web services. It is possible to set up a Meta4 database in less than 30 min on a simple Linux server such as an Amazon EC2 micro instance. Meta4 is a lightweight tool, completely open-source, easy to install locally and easy to add additional functionality through web services.

[3]https://github.com/MG-RAST/

[4]http://blog.metagenomics.anl.gov/mg-rast-v3-2-faq/#local_install

Meta4 was developed on an Amazon EC2 micro instance using a CloudBioLinux (Afgan et al., 2012) image. All code is available via Github. An example Meta4 database can be queried at http://www.ark-genomics.org/tools/meta4 containing an assembly of the Hess et al. (2011) data.

MATERIALS AND METHODS

The overall structure of Meta4 is shown in **Figure 1**. Central to the system is the Meta4 MySQL database, which stores information on samples, assemblies, gene predictions, and protein domain information. The choice to store some basic annotation in the database itself allows users to query the available gene predictions on domains of interest. Without such annotation, it would be very difficult for users to filter the large numbers of gene predictions in metagenomic datasets. We have chosen to store information on protein domains, rather than the results of homology searches (e.g., BLAST), as often domain searches are more sensitive to distant homology. Information can be loaded into the database from common formats using the database loading scripts, including GFF3 (gene predictions) and fasta (contigs and scaffolds). A web form is provided that allows users to query the database and information is presented in two ways: firstly, data extracted directly from the Meta4 database is presented in the browser; secondly, data extracted from the Meta4 database is provided to a range of web services, and the results of those web services presented in the browser. This allows for the latest, live, up-to-date annotation to be displayed for each gene prediction, and is a key feature of Meta4.

INTERFACE AND WEB SERVICES

The dynamic web interface is written in Perl/CGI and should run on any apache web-server with minimal setup. The user is presented with a form including several parameters for search and retrieval of genes/proteins within the database. The results are

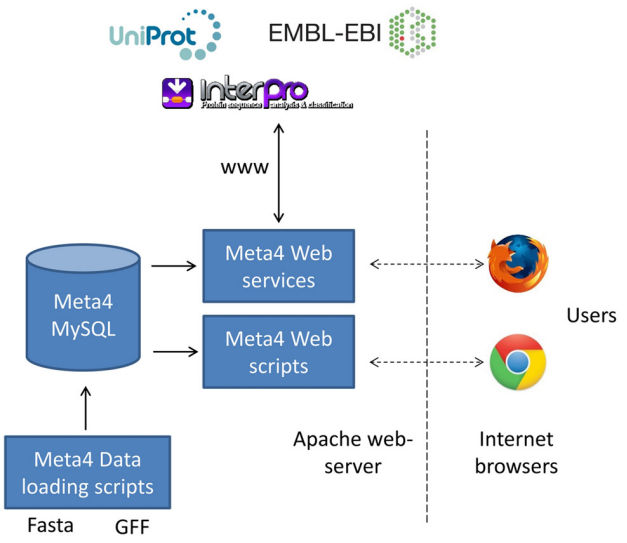


FIGURE 1 | The overall structure of Meta4, which shows the relationship between the MySQL database, the data loading scripts, the web interface, external web services, and the users.

returned as an HTML table, and consist of two parts – those that return information stored in the database, and those returned from web services.

We have implemented three web services in Meta4. The first uses the EBI's SOAP wublast interface (McWilliam et al., 2009), querying Uniprot (Magrane and Consortium, 2011) with a protein sequence retrieved from the database. The top 10 results are returned and these represent the most up-to-date homology information for that protein within Uniprot.

The second uses the Uniprot REST web service (Jain et al., 2009). Domains associated with a particular protein are extracted from the database and used as input to search Uniprot. In this way, known proteins with a similar domain structure to that being queried are returned and presented to the user. Users are then able to see the protein name and species of similar proteins, and can click through to the Uniprot entry.

The third uses the EBI's InterproScan (Mulder and Apweiler, 2007) SOAP interface (McWilliam et al., 2009), querying up to 14 separate protein domain databases with a protein sequence retrieved from the database. The image and text returned also represent the most up-to-date information publicly available for the domains predicted within the query protein.

DATABASE STRUCTURE

The Meta4 MySQL database models the following specific entities and their relationships:

(i) Sample: information about a specific biological sample that has been sequenced. In reality we imagine most researchers will store this information in some other database [e.g., a laboratory information management system (LIMS)], but this table allows metagenomic data to be linked to specific samples.
(ii) Assembly: information about a *de novo* assembly of data from a biological sample. This allows for multiple different assemblies of the same sample. The parameters of the assembly can be stored as tag = value pairs in an assembly_param table.
(iii) Contig: models the contigs that are output as the result of an assembly. We do not explicitly differentiate between contigs and scaffolds. In this instance, a contig simply describes a single, contiguous sequence obtained from a metagenomic assembly.
(iv) Gene prediction: information on the genes predicted on any given contig, including the location on the contig, and the DNA and protein sequence.
(v) Domain database: contains information on the domain database used and allows each gene prediction to have hits to multiple domain databases [e.g., PROSITE (Sigrist et al., 2010) and Pfam (Punta et al., 2012)] or multiple versions of the same domain database.
(vi) Protein domain: information on the domains within each domain database.
(vii) Domain match: storage of the link between gene predictions and protein domains, including location of the match, bit score and e-value.

Crucially, this structure allows multiple assemblies of the same biological sample, as it is common to carry out multiple genome assemblies using different software and parameter sets (which can be flexibly stored in the assembly_param table). Domain matches from multiple databases may also be stored.

CODE STRUCTURE AND DEVELOPMENT

We have implemented the Meta4 data model in MySQL with an interface written in Perl and Perl CGI. The code has been tested on CloudBioLinux (Afgan et al., 2012) and a local Scientific Linux server, and should work on any standard LAMP server. The github repository contains the following folders:

(i) sql: SQL for creating the MySQL database.
(ii) examples: example files used to create a simple instance of Meta4.
(iii) scripts: perl scripts to load information and data into a Meta4 database.
(iv) cgi_scripts: perl CGI scripts that provide an interface to query the data within a Meta4 database.

A README file is included in the distribution which gives accurate instructions on how to create a Meta4 database that is accessible via a web browser. If the import scripts are run with no parameters, simple instructions are printed to the terminal.

Meta4 is released under an open-source license and we welcome active participation in the project. Whilst Meta4 is suitable for release and publication in its current form, there are many ways in which Meta4 could be developed. For example, currently users must import data using Linux command-line scripts, rather than a graphical user interface (GUI); also, we present scripts to import data from the output of pfam_scan.pl[5], and we welcome contributions that are able to import data from other software formats.

RESULTS

EXAMPLE DATASET

We have created an example Meta4 database and the results can be browsed at http://www.ark-genomics.org/tools/meta4. Briefly, we downloaded data from Hess et al. (2011) (SRA accession SRA023560) and assembled the reads using SOAPdenovo (Li et al., 2010). Open-reading frames greater than 200 bp in length were extracted as putative genes. Pfam-A domains were annotated using pfam_scan.pl[5]. As the experiment was designed to find novel biomass degrading genes, we encourage users to enter "glyco_hydro" into the "Name" field and click "Submit."

BROWSING GENE PREDICTIONS

Meta4 allows users to browse information on particular gene predictions. An example screenshot of such information can be seen in **Figure 2**. Basic information such as the gene name, description, and sequence lengths are extracted from the database. Protein domains annotated within the database are also extracted, and presented as both a table and an image. Furthermore, the actual gene and protein sequences are presented, and formatted correctly. Afterward, live information is presented from the three web services. Firstly, proteins with the same domain structure are extracted from Uniprot, and presented as a table. Secondly, the top 10 BLAST hits against Uniprot/TREMBL are presented. In this way, users are able to see similar proteins in Unprot by domain

[5]ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/PfamScan.tar.gz

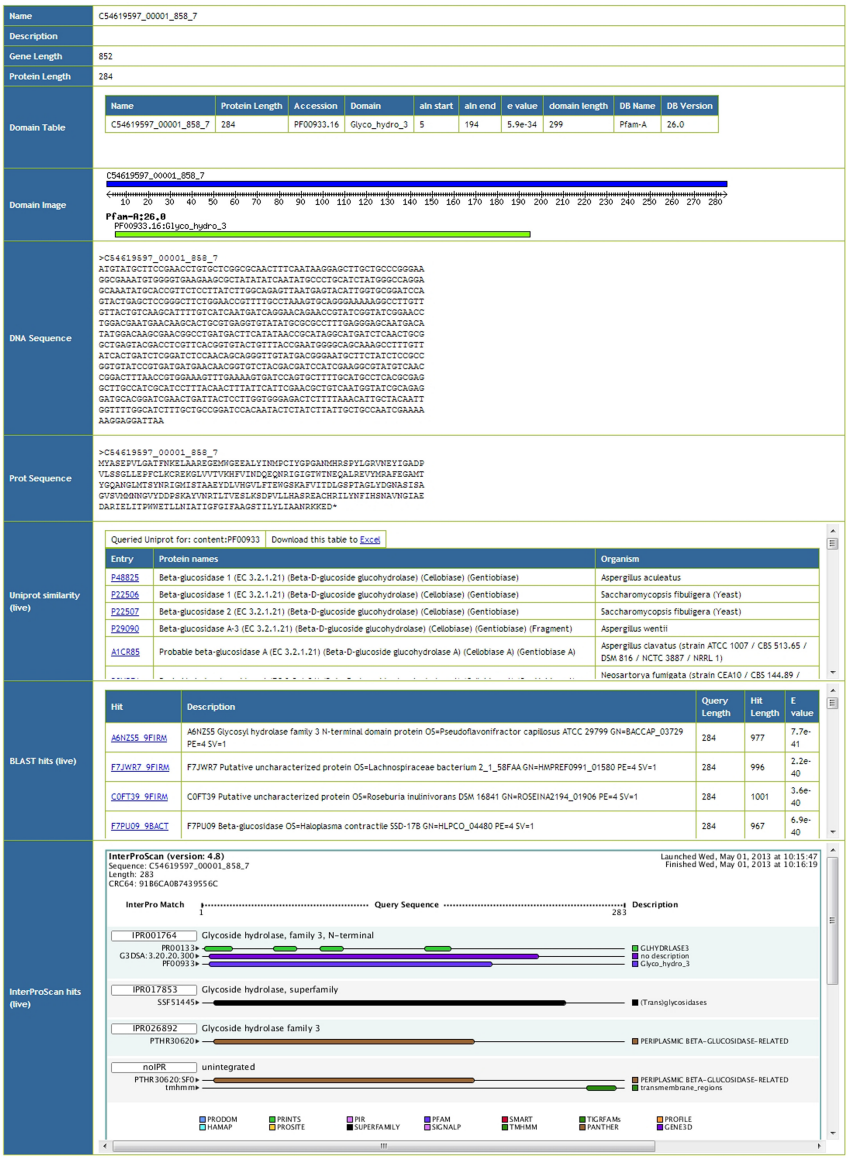


FIGURE 2 | Screenshot of the Meta4 results interface, showing information extracted from the Meta4 database, and information from web services (marked as "live" in the table).

structure and by sequence homology, and can click through to the relevant entries. Finally, results from the InterProScan web service are presented, both as an image and as text. As InterProScan searches 14 different domain databases, we are able to view more information here than the simple domain information stored in the Meta4 database. A key advantage of Meta4 is that information and annotation about the protein in question is served to the user in real time, and therefore represents the most up-to-date information possible.

WEB INTERFACE

The web interface has been tested on Firefox (Windows, Linux, Android), Safari (Windows, Mac), Opera (Windows, Android), Konqueror (Linux), Chrome (Windows), the Android native browser, and Internet Explorer (Windows). All features work on all browsers, except Internet Explorer 8 (Windows). Our implementation of the EBI's InterproScan web service produces an in-line image using the data URI (uniform resource identifier) scheme, and we understand Internet Explorer 8 to have a 32 Kb limit for these. This is fixed in Internet Explorer version 9.

AMAZON EC2 CLOUD IMAGE

An Amazon Machine Image (AMI) is available (EU-WEST: ami-46687f32). The AMI is based on Ubuntu Precise 12.04 (64 Bit) with additional dependencies installed, including Meta4. We have loaded the example data packaged with Meta4, and the system is available from the cgi-bin of the installed Apache2

web-server. Full instructions on how this was set up are available here: http://www.ark-genomics.org/services-bioinformatics-meta4/creating-meta4-amazon-machine-image-ami

DISCUSSION

The role of Meta4 is to allow bioinformaticians to share the results of metagenomic assembly and annotation with collaborators, and to provide those collaborators with a simple web-based interface with which to query and browse the data. It is not intended to compete with tools that aim to assemble, annotate, and functionally or taxonomically compare multiple metagenomic datasets; rather, it is a simple web application that can be used to search and browse large amounts of information quickly, and retrieve genes and proteins that may be of interest for further studies.

The key advantages of Meta4 are:

(i) Simplicity: Meta4 is incredibly simple and can be installed in minutes on a standard LAMP server, either using the git repository or by using the Amazon EC2 image. A new Meta4 instance can be created rapidly from standard formats using the scripts provided. In addition, Meta4 is completely open-source.

(ii) Use of web services: by using web services, Meta4 ensures the latest annotation results are delivered to users. In contrast, other systems store pre-computed results which can rapidly become out-of-date. By using web services, it is easy to extend the functionality of Meta4.

(iii) Separation of data delivery from data analysis: existing web-based systems combine assembly and annotation with results presentation. By separating the search/browse function from data analysis, Meta4 allows bioinformaticians to use an assembly and annotation pipeline of their choice, and still share their results with collaborators through a user-friendly web interface.

(iv) Access control: often when one submits data to a public web-server, a commitment is made to make the data publicly available. Meta4 can be set up on a private intranet in minutes, ensuring data privacy; alternatively, cloud Meta4 instances can be limited to specific IP addresses. Thus Meta4 allows both public and private sharing of data.

Managing the large amounts of data from WGS metagenomics projects is a challenge and there is a need for simple tools that enable scientists to access and query the results. We present Meta4, a simple database for the storage of proteins and their domains predicted from metagenomics experiments. Meta4 is lightweight, easy to install and deploy, and can handle large amounts of data. The system presents information to scientists in a format they understand via a web interface. Meta4 is easily extensible through the addition of web services, and despite not being as feature-rich as some existing systems, benefits from being open-source, lightweight and easy to install and deploy. The use of web services means that the data served to users is as up-to-date as the underlying primary database, which is an advantage over large data warehouses whose data may become out-of-sync with the primary data source. Meta4 is available under an open-source license at http://www.ark-genomics.org/bioinformatics/meta4.

Despite the increasing number of published algorithms for metagenomic assembly and annotation, the complexity of the problem is such that errors are common. Attempts must be made to assess the quality of metagenomic assemblies prior to annotation, especially to ensure inappropriate joins are not made during the contig and scaffold production steps. Metagenomic assemblies are often highly fragmented, and this can affect gene prediction and protein domain annotation. Once specific protein targets have been identified from metagenomic datasets, we recommend a manual annotation step to ensure the gene location (start and end) and protein domain structures are correctly defined.

ACKNOWLEDGMENTS

This research was supported by the Biotechnology and Biological Sciences Research Council (BBSRC; BB/J004243/1, BB/J004235/1), and by the Technology Strategy Board (TS/J000108/1, TS/J000116/1).

REFERENCES

Afgan, E., Chapman, B., Jadan, M., Franke, V., and Taylor, J. (2012). Using cloud computing infrastructure with CloudBioLinux, CloudMan, and Galaxy. *Curr. Protoc. Bioinformatics* Chapter 11, Unit11. 9. doi: 10.1002/0471250953.bi1109s38

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 13, R122. doi: 10.1186/gb-2012-13-12-r122

Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R., and Wittwer, P. (2005). Metagenomic gene discovery: past, present and future. *Trends Biotechnol.* 23, 321–329. doi: 10.1016/j.tibtech.2005.04.001

Cowan, D. A., Arslanoglu, A., Burton, S. G., Baker, G. C., Cameron, R. A., Smith, J. J., et al. (2004). Metagenomics, gene discovery and the ideal biocatalyst. *Biochem. Soc. Trans.* 32, 298–302. doi: 10.1042/BST0320298

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform.* 23, 205–211. doi: 10.1142/9781848165632_0019

Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., et al. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 331, 463–467. doi: 10.1126/science.1200387

Hoff, K. J., Lingner, T., Meinicke, P., and Tech, M. (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* 37, W101–W105. doi: 10.1093/nar/gkp327

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., et al. (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 10:136. doi: 10.1186/1471-2105-10-136

Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., and Salzberg, S. L. (2012). Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* 40, e9. doi: 10.1093/nar/gkr1067

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20, 265–272. doi: 10.1101/gr.097261.109

Mackelprang, R., Waldrop, M. P., Deangelis, K. M., David, M. M., Chavarria, K. L., Blazewicz, S. J., et al. (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480, 368–371. doi: 10.1038/nature10576

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011, bar009. doi: 10.1093/database/bar009

Markowitz, V. M., Chen, I. M., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., et al. (2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.* 40, D123–D129. doi: 10.1093/nar/gkr975

McWilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., Martin, J., et al. (2009). Web services at the European Bioinformatics Institute-2009. *Nucleic Acids*

Res. 37, W6–W10. doi: 10.1093/nar/gkp302

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.* 396, 59–70. doi: 10.1007/978-1-59745-515-2_5

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40, e155. doi: 10.1093/nar/gks678

Noguchi, H., Taniguchi, T., and Itoh, T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 15, 387–396. doi: 10.1093/dnares/dsn027

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M., and Brown, C. T. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc. Natl. Acad. Sci. U.S.A.* 109, 13272–13277. doi: 10.1073/pnas.1121464109

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 27, i94–i101. doi: 10.1093/bioinformatics/btr216

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38, e191. doi: 10.1093/nar/gkq747

Richardson, E. J., and Watson, M. (2013). The automatic annotation of bacterial genomes. *Brief Bioinform.* 14, 1–12. doi: 10.1093/bib/bbs007

Sigrist, C. J., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., et al. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* 38, D161–D166. doi: 10.1093/nar/gkp885

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., et al. (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res.* 39, D546–D551. doi: 10.1093/nar/gkq1102

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., et al. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14, R2. doi: 10.1186/gb-2013-14-1-r2

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74. doi: 10.1126/science.1093857

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Yok, N. G., and Rosen, G. L. (2011). Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics* 12:20. doi: 10.1186/1471-2105-12-20

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 May 2013; accepted: 13 August 2013; published online: 05 September 2013.

*Citation: Richardson EJ, Escalettes F, Fotheringham I, Wallace RJ and Watson M (2013) Meta4: a web application for sharing and annotating metagenomic gene predictions using web services. Front. Genet. **4**:168. doi: 10.3389/fgene.2013.00168*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.

Software

DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data

Michael Watson*, Juliet Dukes†, Abu-Bakr Abu-Median*, Donald P King† and Paul Britton*

Addresses: *Institute for Animal Health, Compton, Newbury, Berks RG20 7NN, UK. †Institute for Animal Health, Pirbright, Surrey GU24 0NF, UK.

Correspondence: Michael Watson. Email: michael.watson@bbsrc.ac.uk

Published: 14 September 2007

Genome ***Biology*** 2007, **8:**R190 (doi:10.1186/gb-2007-8-9-r190)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/9/R190

Received: 1 June 2007
Revised: 15 August 2007
Accepted: 14 September 2007

Abstract

DNA microarrays offer the possibility of testing for the presence of thousands of micro-organisms in a single experiment. However, there is a lack of reliable bioinformatics tools for the analysis of such data. We have developed DetectiV, a package for the statistical software R. DetectiV offers powerful yet simple visualization, normalization and significance testing tools. We show that DetectiV performs better than previously published software on a large, publicly available dataset.

Rationale

One of the key applications of metagenomics is the identification and quantification of species within a clinical or environmental sample. Microarrays are particularly attractive for the recognition of pathogens in clinical material since current diagnostic assays are typically restricted to the detection of single targets by real-time PCR or immunological assays. Furthermore, molecular characterization and phylogenetic analysis of these signatures can require downstream sequencing of genomic regions. Many microarrays have already been produced with the aim of characterizing the spectrum of micro-organisms present in a sample, including detection of known viruses [1-5], assessment of bioterrorism [6,7] and monitoring food quality [8].

However, the use of DNA microarrays for routine applications produces many challenges for bioinformatics. Firstly, probe selection is a difficult and time consuming process. There are a huge number of diverse species in nature, of which we have sequence information for only a tiny fraction. This makes it difficult to find oligonucleotides, either alone or in combination, that uniquely identify species of interest. Oligos may have homology to multiple species, which results in a complex and noisy hybridization pattern. Secondly, each nucleic acid sample tested will typically contain a mixture of DNA and RNA from the organism of interest, the host and from a variety of contaminants, which may all contribute to the resulting microarray profile. Furthermore, this may be complicated by the presence of multiple, possibly related, pathogen species, making it difficult to separate patterns due to cross-hybridization from a true positive result.

Urisman *et al.* [9] have previously reported E-Predict, a computational strategy for species identification based on observed microarray hybridization patterns. E-Predict uses a matrix of theoretical hybridization energy profiles calculated by BLAST-ing completely sequenced viral genomes against the oligos on their array, and calculating a free energy of hybridization. Observed hybridization profiles are then compared to the theoretical profiles using a similarity metric, and a p value calculated using a set of experimentally obtained null probability distributions. E-Predict has been shown to

produce useful results in a number of situations. However, at present, E-Predict does not contain any tools for visualization, and requires extensive customization and calculation before it is applicable to new arrays. Also, E-Predict is only available as a CGI script for Unix/Linux platforms.

We present DetectiV, a package for R [10] containing functions for visualization, normalization and significance testing of pathogen detection microarray data. R is a freely available statistical software package available for Windows, Unix/Linux and MacOS, meaning DetectiV is a platform independent solution. DetectiV uses simple and established methods for visualization, normalization and significance testing. When applied to a publicly available microarray dataset, DetectiV produces the correct result in 55 out of 56 arrays tested, an improvement on previously published methods. When applied to a second dataset, DetectiV produces the correct result in 12 out of 12 arrays.

Implementation

DetectiV is implemented as a package for R, a powerful, open-source software package for statistical programming [10]. Many packages for R already exist for the analysis of biological datasets, including microarray data, and the bioconductor project [11] is just one example of a group of such packages. As it is implemented in R, DetectiV easily integrates with many of the packages available for microarray analysis, including limma [12], marray [11] and affy [13].

DetectiV is written in the native R language and uses standard functions within R. As R is available on Microsoft Windows, Unix (including linux) and MacOS, DetectiV represents a platform independent solution for the analysis of pathogen-detection microarray data.

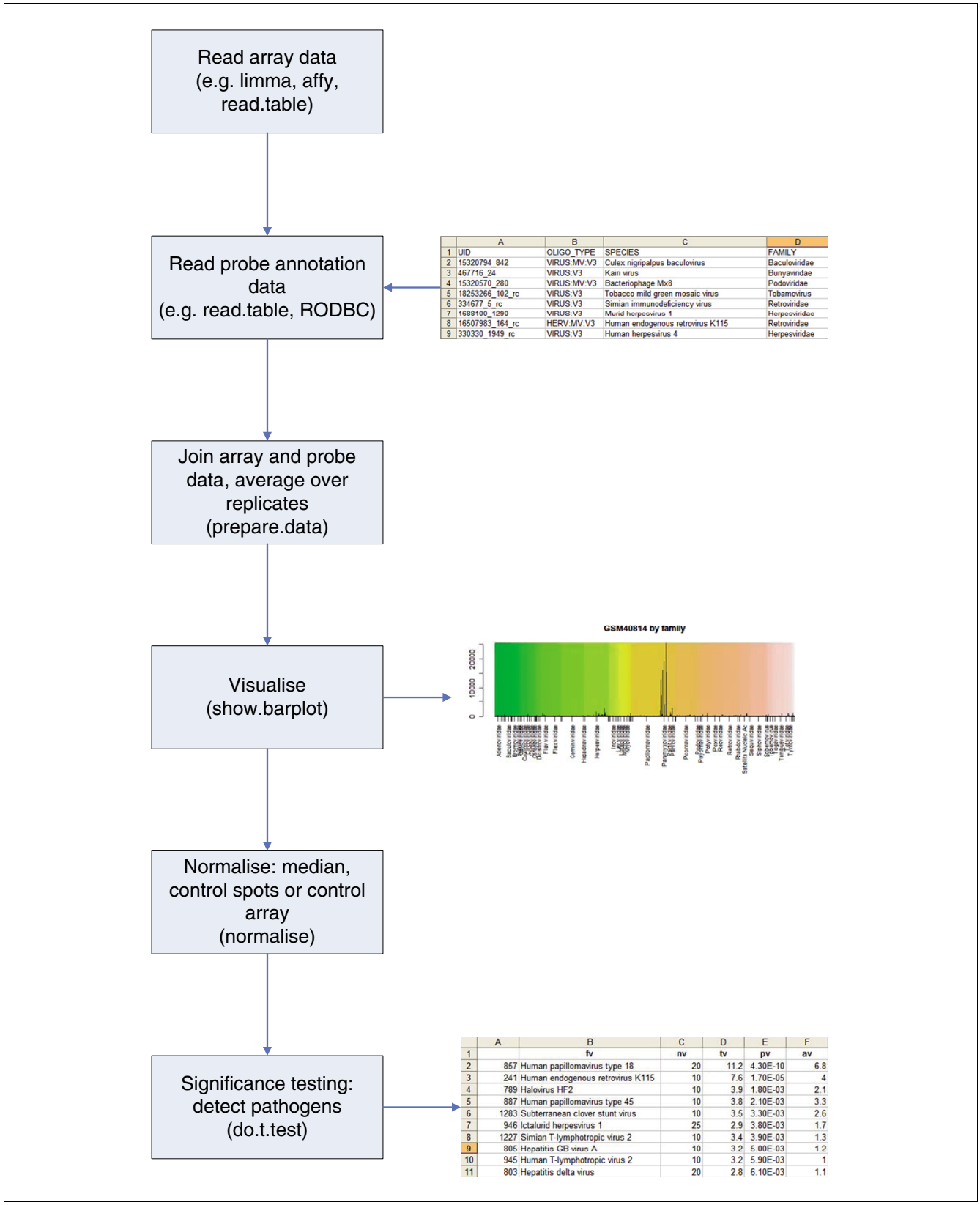
The flow of information through DetectiV is shown in Figure 1. The basic dataset required is a matrix of data, with rows representing probes on the array, and columns representing measurements from individual microarrays. This dataset is easily produced from data structures created by limma [12], which includes functions for reading in many common microarray scanner output formats, and affy [13], which provides functions for reading in affymetrix data. Commonly, researchers will have an additional file of information giving details about each probe. In the case of pathogen detection arrays, this file will most often contain the type, species, genus and other classification data for the pathogen to which each probe is designed. It should be noted that there may be more than one entry in this file for each probe; for example, if a given probe is thought to hybridize to multiple pathogens. In text format, these may be read in using the native read.table command, or in excel format using the RODBC library.

Once these two datasets are in R, DetectiV prepares them for analysis using the prepare.data function. This function joins the array data to the probe information data based on a unique ID. The researcher may choose to subtract local background if appropriate. The default at this stage is to average over replicate probes, again based on a unique ID. This will result in a single value for each unique probe for each array. The data will have one or more columns of extra information from the annotation file, and these columns will be used to group the data for further analysis.

Researchers will wish to visualize their data in order to compare the hybridization signals for the probes recognizing the different pathogen signatures. DetectiV provides a function called show.barplot for this. The output from prepare.data is passed to the function, along with the name of the column containing the variable by which the data will be grouped, referred to here as *group*. An example in pathogen detection data may be species, genus, family, and so on. The data are sorted into unique groups as defined by the unique values of *group*. A barplot is drawn, with one bar per unique probe. Probes from the same group are drawn together. Each group is represented by a unique background color, enabling the user to easily visualize the different groups. An example output is shown in Figure 2. This sample comes from Urisman *et al* [9] and represents data from a virus detection microarray hybridized with amplified RNA from nasal lavage, positive for respiratory syncytial virus by direct fluorescent antibody (DFA) test. The *group* chosen here is virus family. It is quite clear from this image that there is a virus from the family *Paramyxoviridae* present in the sample, demonstrated by the high bars associated with that family.

These images are often very large, and so DetectiV offers the ability to subset the data before plotting by using the get.subset function. Figure 3 shows a similar barplot using a subset of the data: only those oligos representing species that belong to the *Paramyxoviridae* family. It is clear from this image that those oligos representing different groups/species of respiratory syncytial virus have the highest intensity, as we would expect, although there is cross-hybridization with oligos for human metapneumovirus (another paramyxovirus in the same sub-family: Pneumovirinae).

DetectiV may also carry out normalization and significance testing. For this, there is the function normalise. Here, the aim of normalization is to represent the data in relation to a negative control. The idea is that if the values for each probe are divided by the negative control and then the log2 taken, then the data should be normally distributed, and each group should have a mean of zero (providing a pathogen is not present). Traditional statistical tests can then be used to test if any group of probes is significantly different from zero. DetectiV offers three methods of normalization, each using a different 'type' of negative control, and these are summarized in Table 1.



	A	B	C	D
1	UID	OLIGO_TYPE	SPECIES	FAMILY
2	15320794_842	VIRUS:MV:V3	Culex nigripalpus baculovirus	Baculoviridae
3	467716_24	VIRUS:V3	Kairi virus	Bunyaviridae
4	15320570_280	VIRUS:MV:V3	Bacteriophage Mx8	Podoviridae
5	18253266_102_rc	VIRUS:V3	Tobacco mild green mosaic virus	Tobamovirus
6	334677_5_rc	VIRUS:V3	Simian immunodeficiency virus	Retroviridae
7	1698100_1290	VIRUS:V3	Murid herpesvirus 1	Herpesviridae
8	16507983_164_rc	HERV:MV:V3	Human endogenous retrovirus K115	Retroviridae
9	330330_1949_rc	VIRUS:V3	Human herpesvirus 4	Herpesviridae

	A	B	C	D	E	F
1		fv	nv	tv	pv	av
2	857	Human papillomavirus type 18	20	11.2	4.30E-10	6.8
3	241	Human endogenous retrovirus K115	10	7.6	1.70E-05	4
4	789	Halovirus HF2	10	3.9	1.80E-03	2.1
5	887	Human papillomavirus type 45	10	3.8	2.10E-03	3.3
6	1283	Subterranean clover stunt virus	10	3.5	3.30E-03	2.6
7	946	Ictalurid herpesvirus 1	25	2.9	3.80E-03	1.7
8	1227	Simian T-lymphotropic virus 2	10	3.4	3.90E-03	1.3
9	806	Hepatitis GB virus A	10	3.2	5.00E-03	1.2
10	945	Human T-lymphotropic virus 2	10	3.2	5.90E-03	1
11	803	Hepatitis delta virus	20	2.8	6.10E-03	1.1

Figure 1
Flow of information, and steps taken, when analyzing pathogen detection microarray data using DetectiV.

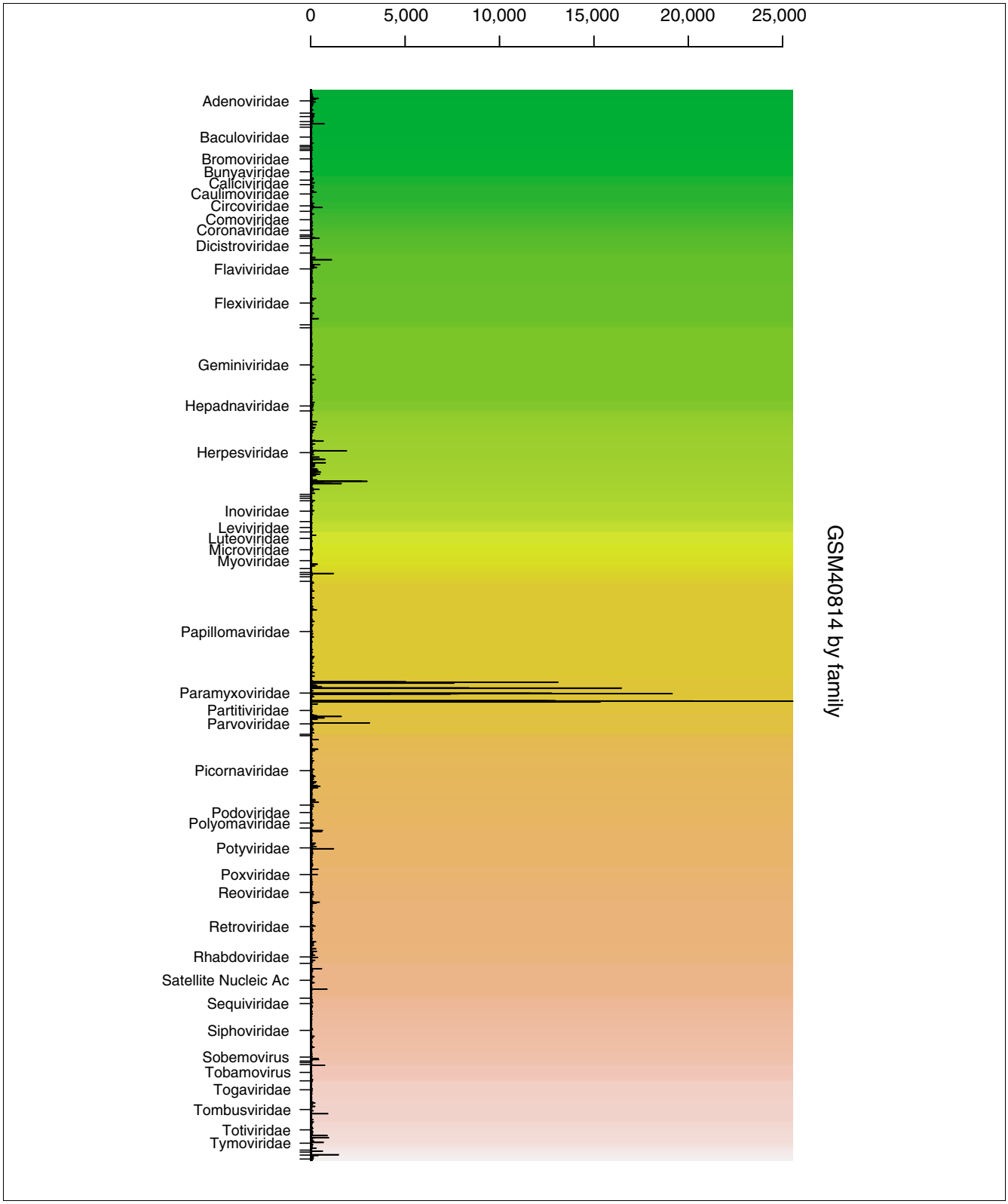


Figure 2
GSM40814 by family. Example barplot from DetectiV showing data from a virus detection microarray. The sample included amplified RNA from nasal lavage, positive for respiratory syncytial virus by DFA. Oligos have been averaged over replicates and grouped according to virus family. Each unique oligo is represented by a single bar. Each virus family has a unique background color. The y-axis is raw intensity.

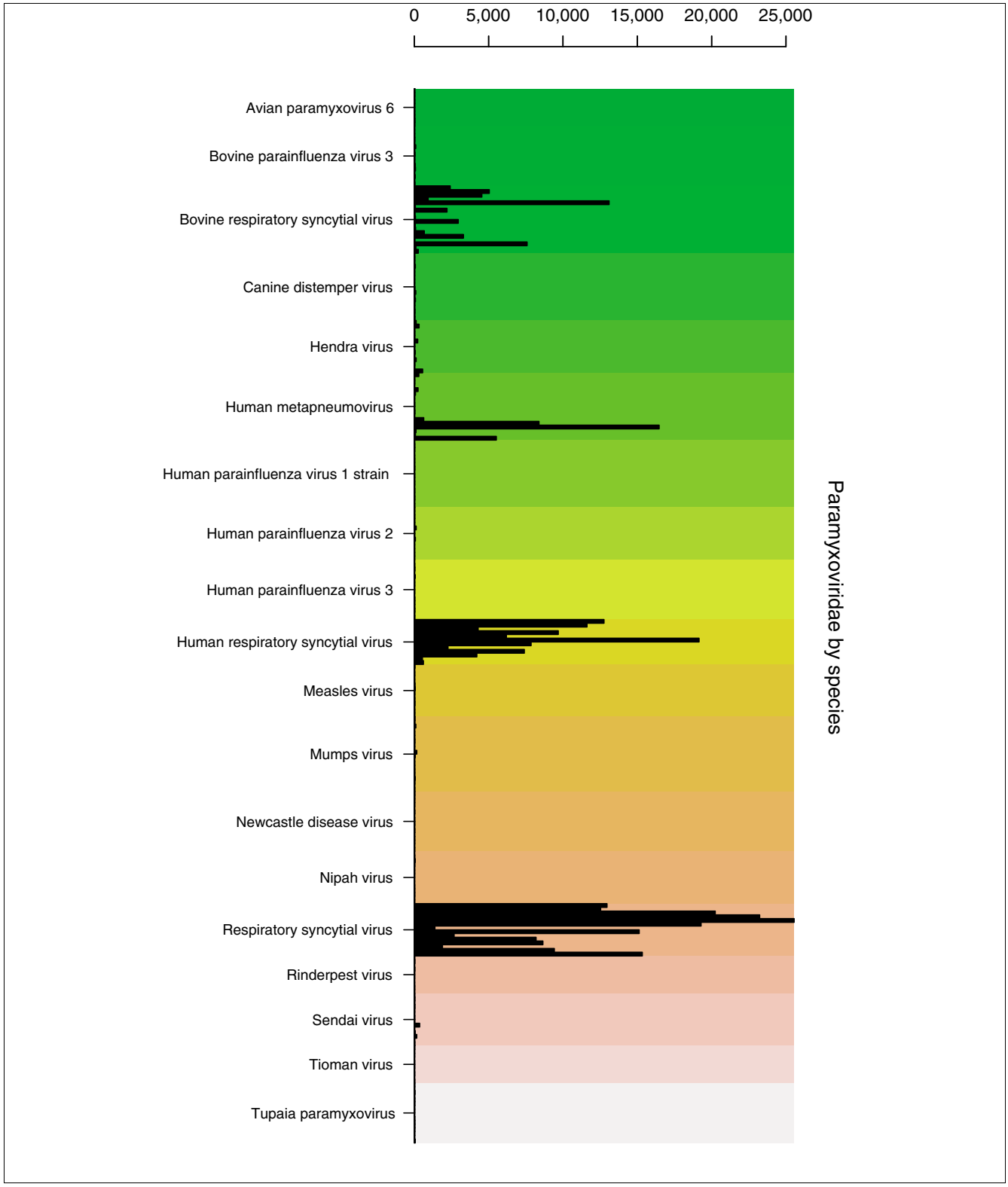


Figure 3
GSM40814 Paramyxoviridae by species. Example barplot from DetectiV showing data from a virus detection microarray. The sample included amplified RNA from nasal lavage, positive for respiratory syncytial virus by DFA. Only oligos representing species from the Paramyxoviridae family are shown. Oligos have been averaged over replicates and grouped according to virus species. Each unique oligo is represented by a single bar. Each virus species has a unique background color. The y-axis is raw intensity.

Table 1

DetectiV normalization methods

Method	Normalized statistic	Terms
Median	$\log_2\left(x_j^i/\tilde{x}_j\right)$	Where x_j^i is the value for probe *i* on array *j* and $\tilde{x}_j$ is the median value for all probes on array *j*
Control	$\log_2\left(x_j^i/\bar{c}_j\right)$	Where x_j^i is the value for probe *i* on array *j* and $\bar{c}_j$ is the mean value for control oligo *c* on array *j*
Array	$\log_2\left(x_j^i/x_c^i\right)$	Where x_j^i is the value for probe *i* on array *j* and x_c^i is the value for probe *i* on control array/channel *c*

Explanation of the three normalized statistics offered by DetectiV.

The median method calculates the global median value for each array. It should be noted that this method assumes that most probes will not hybridize to anything. If this assumption is false then this method should not be used. However, if the assumption holds, then the median is a good representation of that value we would expect to see from probes that have not hybridized to anything.

The control method relies on specific negative controls having been spotted on the array. The researcher may then choose one of these controls, and the mean value is calculated for that control for each of the arrays. The mean control value for each array is then used as a divisor for each probe on their respective arrays.

Finally, the array method utilizes an entire control array or channel. In this instance, an entire array is chosen to be the negative control, and all probe values are divided by their respective elements from the control array. An obvious example for a control array may be RNA from a known uninfected animal. The control array therefore has a value for each specific probe representing that value we would expect to see if that specific probe has not hybridized to anything.

In all instances, after taking the log2, groups of probes that have not hybridized to anything should be normally distributed and have mean zero. We can therefore split the probes into groups and perform a *t*-test for each one. DetectiV does this using the do.t.test function. The normalized (or raw) data are split into groups as defined by the unique values of a user defined annotation column. Providing each group has more than two probes, a *t*-test is performed to test the difference of the observations from zero. The average value is also calculated. The output is a table, sorted by *p* value.

Methods and data analysis

The data used were downloaded from the Gene Expression Omnibus (GEO) [14], accession number GSE2228. The array platform for this data is GEO accession GPL1834, and includes over 11,000 oligos representing over 1,000 viral and bacterial species [4].

The dataset itself consists of 56 arrays including 15 independent HeLa RNA hybridizations, 10 independent nasal lavage samples positive for respiratory syncytial virus, 7 independent nasal lavage samples positive for influenza A virus, a serum sample positive for hepatitis B virus, a nasal lavage sample positive for both influenza A virus and respiratory syncytial virus, and culture samples of 11 distinct human rhinovirus serotypes.

Both DetectiV and E-Predict [9] have been used to analyze the data. For DetectiV, the data were not corrected for local background. Missing, negative and zero values were set to a nominal value of 0.5. Intensities were averaged across replicate probes. Median normalization was then carried out, followed by a *t*-test grouping the data by virus species. Probes representing actin, GAPDH and Line_Sine were filtered from

Table 2

E-Predict parameters

Parameter	Value
user_wts	MV_72worst_medRaw500_badYdens
norm_opt	Sum
energy_filter	undef
ematrix	22/07/2004
ematrix_norm	Quadratic
ematrix_efilter	30
dist_metric	Pearson Uncentered
iterate	2
top_oligos	5
top_genomes	5
top_fams	5
sort_by	Distance|P value
eclust	None

Parameters used for input into E-Predict.

the results. Results were first filtered such that groups had a normalized log2 ratio greater than or equal to 1 (a ratio of two to the control) and then sorted by *p* value. This method will be referred to as DetectiV.

For E-Predict, default values for all parameters were used, and are shown in Table 2. Data points were corrected for local background, as per the examples in Urisman *et al.* [9]. E-Predict filters out 266 oligos by default, and this setting was kept. In all cases, E-Predict carried out two iterations, although only results from the first iteration are shown here. The best performing method of interpreting the results was to take those species with a p value ≤ 0.05 and sort by distance (termed E-Predict.dist). Note that this is the method cited in [9], example 3, used to demonstrate E-Predict's ability to detect SARS.

Pathogen detection arrays have also been implicated in the discovery of SARS. Urisman *et al.* [9] reported that although their original platform did not contain oligos designed to SARS, once the SARS genome had been published, it was possible to recalculate the energy matrix for E-Predict and find that the energy profile for SARS was the top hit (after taking those viruses with low *p* values and sorting by distance). We have applied DetectiV to the same dataset (GEO accession GSE546). To include oligos for SARS, we searched a database of oligo sequences on the array with sequence NC_004718 from RefSeq using NCBI blast. There were 61 oligos on the array that hit the SARS genome with greater than 80% identity across an alignment of 20 bp or more. In the analysis, these oligos were assigned as representative of two viruses: their original virus and SARS. The data were median normalized and a *t*-test carried out using DetectiV.

Finally, having established that DetectiV compares favorably with previously published software, we have validated the DetectiV software by applying it to a second dataset. The data used were downloaded from the GEO [14], accession number GSE8746. The array platform for this data is GEO accession GPL5725, and consists of 5,824 oligos representing over 100 viral families, species and subtypes. The dataset itself consists of 12 arrays, 4 hybridized with RNA from cell cultured foot-and-mouth disease virus (FMDV) type O, 3 hybridized with RNA from FMDV type A, 1 hybridized with RNA from a sheep infected with FMDV type O, and 4 hybridized with cell-cultured avian infectious bronchitis virus (IBV). Analysis using DetectiV was carried out as described above.

Results and comparison

We present here results from two methods of analysis, termed DetectiV and E-Predict.dist, as described above. There are 56 arrays in the dataset, the expected results of which are known. Each array was hybridized with RNA containing a single virus, except GSM40845, which was infected with both influenza A and respiratory syncytial virus. We assigned a correct result for each method if the top hit from the analysis was the same as the known infectious agent or, if that agent was not represented on the array, the top hit was a very closely related virus. In the case of GSM40845, we report a correct result if both viruses were at the top of the reported hits, to the exclusion of other virus species (but not closely related strains).

Additional data file 1 gives the top hit for both analysis methods in all 56 arrays. As can be seen, DetectiV generated a correct result in 55 out of the 56 arrays. In comparison, the E-Predict.dist method gave a correct result in 53 out of the 56 arrays. These results are discussed in greater detail below.

DetectiV

Full results for each of the arrays can be found on the DetectiV website [15]. Within the 55 correct results, there are three classes that require slightly different interpretation, examples of which are GSM40806, GSM40810 and GSM40817. Results for these arrays are given in Table 3.

Array GSM40806 was hybridized with amplified HeLa RNA, and the top hit from DetectiV is human papillomavirus type 18, as expected. This virus has both the smallest *p* value and largest mean normalized log ratio. There is also clear

Table 3

Typical results from DetectiV

GSM40806			GSM40810			GSM40820		
Virus	*p* value	Mean	Virus	*p* value	Mean	Virus	*p* value	Mean
Human papillomavirus type 18	4.1E-10	6.8	Human rhinovirus sp.	9.9E-12	4.1	Human herpesvirus 5	5.3E-16	0.57
Human endogenous retrovirus K115	0.000016	4	Human rhinovirus A	2.3E-09	4.1	Respiratory syncytial virus	1.1E-09	4.26
Halovirus HF2	0.0017	2.1	Enterobacteria phage M13	2.2E-07	5.7	Human rhinovirus sp.	5.9E-08	0.75
Human papillomavirus type 45	0.002	3.3	Human rhinovirus 16	6.2E-07	3.5	Human rhinovirus B	1.4E-07	0.47
Subterranean clover stunt virus	0.0032	2.6	Human rhinovirus 1B	0.000001	3.5	Human rhinovirus A	6E-07	0.75

Top five hits from three microarrays showing typical results from DetectiV. All have been sorted by *p* value. GSM40806 and GSM40810 have been filtered such that mean ≥ 1.

distinction between the top hit and the rest of the hits below; there are orders of magnitude between the values for both the *p* value and the mean normalized log ratio. The other hits in the table are expected as a result of hybridization by the virus and host RNA to non-specific probes on the array. However, the clear distinction in both the *p* value and mean log ratio identify human papillomavirus type 18 as the top, and only, hit.

GSM40810 was hybridized with RNA containing human rhinovirus 28. There are 24 distinct groups of human rhinoviruses represented on the array, including a group of oligos for all members ('human rhinovirus sp.), one each for human rhinovirus A and B, and several groups for distinct serotypes. Human rhinovirus 28 is not one of those serotypes specifically targeted by the array; however, as a serotype of the human rhinovirus A species, we would expect the groups for human rhinovirus sp. and human rhinovirus A to be prevalent amongst the results. As can be seen from Table 3, the top hit from DetectiV is human rhinovirus sp., closely followed by human rhinovirus A, the expected result. The reason we have highlighted this array, however, is that the result for Enterobacteria phage M13 shows a higher mean normalized intensity than any of the rhinovirus groups. This is representative of a class of result from DetectiV whereby a virus group has a higher mean normalized log ratio, but a larger *p* value, than the top hit. Here, as in GSM40806, we see orders of magnitude between the *p* value for the top hit and that for Enterobacteria phage M13, which identifies human rhinovirus as being the infectious agent, but in this case we cannot rely on the mean normalized intensity. In this particular instance, Enterobacteria phage M13 is represented by 10 oligos, all of which have intensities far greater than the global median, but which vary considerably between 982 and 18,864. These high values may be due to hybridization with a cloning vector.

Finally, array GSM40817 was hybridized with respiratory syncytial virus. The results are again shown in Table 3, but for this array only, they have not been filtered on mean normalized intensity. Human herpesvirus 5 has by far the smallest *p* value of any of the virus groups; however, it also has a very small mean normalized log ratio. The correct hit, respiratory syncytial virus, has the second smallest *p* value, but has a much larger mean normalized log ratio. This represents the final class of result seen by DetectiV, where the correct virus group does not have the smallest *p* value, but does have a much larger mean normalized log ratio than those groups that have smaller *p* values. The small *p* value of respiratory syncytial virus combined with the large mean normalized log ratio identifies respiratory syncytial virus as the only infectious agent. In this instance, human herpesvirus 5 is represented by 241 oligos, 167 of which are greater than the global median, but all of which have intensities less than 1,000. This could be due to the oligos for human herpesvirus 5 having distant homology with the infectious agent or host cell.

Table 4

Incorrect DetectiV result

Virus	*p* value	Mean
Human herpesvirus 7	8.60E-06	1.7
Bovine respiratory syncytial virus	2.70E-04	2
Respiratory syncytial virus	3.30E-04	3.2
Ictalurid herpesvirus 1	1.50E-03	1.7
Human herpesvirus 6B	1.50E-03	1.8

Top five hits from the DetectiV method from array GSM40816. The sample for this array was found to contain respiratory syncytial virus by DFA.

These three types of result are typical of DetectiV, and explain why both the *p* value and the mean normalized log ratio must be taken into account when interpreting the results. Thus, if the results from DetectiV are filtered such that only viruses whose mean normalized log ratio is ≥ 1, and then sorted by *p* value, the three scenarios described here are accounted for, and we obtain the correct result in 55 out of the 56 arrays.

The single incorrect result for DetectiV comes from GSM40816, which reports human herpesvirus 7 as the top hit, whereas the infectious agent was in fact respiratory syncytial virus. The top five hits for this array using the DetectiV method are shown in Table 4. As can be seen, bovine respiratory syncytial virus and respiratory syncytial virus are second and third, respectively. Both respiratory syncytial virus and bovine respiratory syncytial virus have higher mean values than human herpesvirus 7, although the latter has a smaller *p* value and a mean value that is above the cut-off of 1. Had the results been filtered for p value ≤ 0.5 and then ordered by average value, then the top hit would have been respiratory syncytial virus; similarly, if a cut-off of 2 had been applied instead of 1, a correct result would have been reported. However, across the entire dataset these methods of interpreting the results perform worse than the DetectiV method described above. It is worth noting here that for this array, E-Predict gives the correct top hit.

E-Predict

The results from E-Predict follow similar patterns to those of DetectiV. In most cases it is obvious which virus is the infectious agent, either by examining the *p* value, the similarity or both together. Full results can be seen on the DetectiV website [15]. However, there are certain results reported by E-Predict where it is impossible to obtain the correct result no matter which combination of *p* value and similarity is used. These arrays are arrays are GSM40809, GSM40821 and GSM40847, and the top five results for these arrays can be seen in Table 5.

GSM40809 was hybridized with RNA containing human rhinovirus 26. Again, this is a serotype not specifically targeted by the array; however, as a serotype of human rhinovirus B we

Table 5

Incorrect E-Predict results

GSM40809			GSM40821			GSM40847		
Virus	*p* value	Similarity	Virus	*p* value	Similarity	Virus	*p* value	Similarity
Human enterovirus D	0.000043	0.258894	Orangutan hepadnavirus	0.002291	0.148865	Human enterovirus B	0.000014	0.386095
Human rhinovirus B	0.000045	0.267815	Hepatitis B virus	0.002376	0.147182	Human enterovirus A	0.000016	0.378912
Human enterovirus C	0.000052	0.254504	Woodchuck hepatitis B virus	0.002716	0.10964	Human echovirus 1	0.000022	0.414618
Enterovirus Yanbian 96-83csf	0.000094	0.276873	Woolly monkey hepatitis B Virus	0.00284	0.128919	Enterovirus Yanbian 96-83csf	0.000022	0.412299
Human echovirus 1	0.000134	0.253816	Arctic ground squirrel hepatitis B virus	0.003227	0.103357	Human enterovirus D	0.000026	0.296065

Top five results from the E-Predict.dist method for arrays GSM40809, GSM40821 and GSM40847. In all cases results are ordered by *p* value.

would expect the 'human rhinovirus sp.' and 'human rhinovirus B' groups to be the top hits (this is the case for DetectiV). However, E-Predict reports human enterovirus D as having the smallest *p* value, and enterovirus Yanbian 96-83csf as having the largest similarity. The top five hits reported in Table 5 for this array all have similar *p* values and similarity measures, and there is no way of sorting or filtering the results such that human rhinovirus B becomes the top hit. Without the *a priori* knowledge that human rhinovirus 26 was the infectious agent, it would be more likely to conclude that a species of enterovirus was present in the sample. It is no surprise that these viruses are being confused, as they are related viruses from the Picornaviridae family. However, DetectiV is capable of calling the correct result in this instance, whereas E-Predict is not.

Array GSM40821 was infected with hepatitis B virus but E-Predict reports orangutan hednavirus as having both a smaller *p* value and a higher similarity. This is not that surprising given that hepatitis B and orangutan hepadnavirus are closely related; however, the fact remains that with no *a priori* knowledge, the only logical conclusion from this result would be that the infectious agent was orangutan hepadnavirus. Again, DetectiV calls this array correctly.

Finally, array GSM40847 was hybridized with RNA containing human rhinovirus 87. Again, this is a serotype not specifically targeted by the array, and is not present in the NCBI taxonomy database [16] at the time of writing. We can therefore expect the 'human rhinovirus sp.' group to be high amongst the results (in fact, it is the top result for DetectiV). E-Predict reports human enterovirus B as having the smallest *p* value and human echovirus 1 as having the largest similarity. In fact, E-Predict does not report any rhinovirus oligos in the first iteration at all, and it is only in the second iteration that the group human rhinovirus A is reported as significant.

In the three cases outlined above, there is no clear way of distinguishing the incorrect virus from the correct one. There is also no consistent method of sorting or filtering the results that would give the correct results. In these three cases, E-Predict is unable to distinguish closely related virus species and serotypes. We have reported here the best performing method of interpreting E-Predict results, whereby virus groups with a p value ≤ 0.05 are sorted by distance. This results in a success rate of 53 out of 56 arrays.

DetectiV and SARS

The top five hits from the analysis of the SARS dataset can be found in Table 6. As can be seen, the top hit is SARS, with the lowest *p* value and the highest mean normalized log ratio. SARS is distinct from the other viruses, having a *p* value three orders of magnitude lower than the second top hit.

Validation

Full results can be found on the DetectiV website [17]. The top hit from DetectiV for each of the 12 arrays from GSE8746 can be found in Table 7. As can be seen, DetectiV clearly identifies the infectious agent in all 12 cases. DetectiV works for both the cell-cultured samples and the infected sheep, and shows the ability of the array to distinguish between different subtypes of FMDV.

Discussion

Developing a quick and reliable test for the presence/absence of thousands of bacterial and viral species in a single experiment is an attractive proposition, and a function that DNA microarrays are ideally suited to. Microarrays are extremely high-throughput and relatively cheap. In the case of pathogen detection, the aim must be to quickly and clearly identify those pathogens present in a sample with high confidence, keeping false positives and false negatives to a minimum.

However, the data from such microarrays pose many problems. Firstly, oligos may not be unique to the species they are designed to. For certain species it is impossible to find a large number of oligos that are unique only to that virus that meet the criteria for oligo selection. This is particularly problematic

Table 6

DetectiV results for SARS array

Virus	*p* value	Mean
SARS	8.43E-09	1.906095
Human herpesvirus 7	3.29E-06	1.292008
Simian retrovirus 2	4.27E-05	1.328653
Coliphage alpha3	6.08E-05	1.113462
Transmissible gastroenteritis virus	7.88E-05	1.463675

Top five results from the DetectiV method of analyzing array GSM8528 from GEO accession GSE546. The sample hybridized to the array contained the SARS virus.

Table 7

Top hit for GSE8746

Array	RNA	Top hit	*p* value	Mean
GSM216542	Amplified RNA from cell cultured FMDV type O	FMDO	1.51E-25	2.296645
GSM217164	Amplified RNA from cell cultured FMDV type O	FMDO	1.07E-45	3.513068
GSM217167	Amplified RNA from cell cultured FMDV type O	FMDO	2.36E-48	3.446262
GSM217169	Amplified RNA from cell cultured FMDV type O	FMDO	5.91E-30	2.827877
GSM217172	Amplified RNA from cell cultured FMDV type A	FMDA	6.96E-30	3.560941
GSM217175	Amplified RNA from cell cultured FMDV type A	FMDA	8.71E-14	1.553392
GSM217177	Amplified RNA from sheep infected with FMDV type O	FMDO	1.12E-27	2.431874
GSM217180	Amplified RNA from cell cultured FMDV type A	FMDA	2.97E-33	3.609092
GSM217183	Amplified RNA from cell cultured Avian IBV	IBV	1.05E-21	5.262134
GSM217184	Amplified RNA from cell cultured Avian IBV	IBV	3.49E-33	7.958662
GSM217186	Amplified RNA from cell cultured Avian IBV	IBV	6.20E-33	7.827526
GSM217188	Amplified RNA from cell cultured Avian IBV	IBV	1.44E-35	8.0118

The top hit from DetectiV for the 12 arrays from the GSE8746 dataset. DetectiV produces the correct result in all 12 cases.

for closely related species and strains. In such cases, the 'best' oligos are added to the array, in the knowledge that multiple viruses may hybridize to them. This leads to noisy signals across multiple virus families, species and serotypes. Secondly, infected biological samples may contain many different virus species and strains, making interpretation difficult. Thirdly, it is known that certain oligos simply do not work, even when the array is hybridized with the species that those oligos were designed to. Without testing the array with each virus, we are incapable at present of predicting which oligos will work and which will not. With thousands of species per array, many of which cannot be cultured *in vitro*, it is unfeasible to challenge arrays with every species. Finally, we of course do not know, nor can we ever know, the complete genome sequence of every virus we may encounter. Therefore, though we think we have oligos unique to a species or strain, that is only ever in the context of our knowledge at the time of design, and they may not in fact be unique.

Despite these problems, many species detection arrays have been developed [1-5]. However, reliable methods of data analysis have been rare. Initial methods included visual inspection of the array [4] and clustering [18], both of which are subjective and time-consuming. To combat this, Urisman *et al.* [9] have proposed a more robust method, E-Predict. E-Predict utilizes a pre-calculated energy matrix for each oligo on the array and uses a variety of normalization and similarity metrics to calculate a *p* value and similarity for each virus. The advantages of E-Predict are that it is quantitative, produces good results and is extensible, through the extension of the energy matrix. The disadvantages of the software are a lack of visualization tools, the need to customize parameters for different array platforms and hybridization conditions, and the availability of the software only as a CGI script on the Unix/Linux platform.

We have developed DetectiV, a package for R containing visualization, normalization and significance testing functions for pathogen detection data. DetectiV uses simple and well established visualization and statistical techniques to analyze data from pathogen detection microarrays. DetectiV offers a powerful visualization option in the form of a barplot, enabling researchers to quickly and easily identify possible infectious agents. Data can then normalized to a negative control

(be that a specific probe, array or the global median), transformed by taking the log2 and then subjected to a *t*-test for each species on the array. Oligos are allowed to represent any number of viruses, and thus any analysis is easily extensible by simply updating the list of which oligos represent which species.

DetectiV requires minimal set up and configuration, requiring only an additional file detailing which species each oligo represents. In the majority of cases, these files will already exist. It is then possible to apply DetectiV 'out of the box' to any array data that is readable by R or bioconductor. DetectiV requires no training, configuration or customization specific to each array. DetectiV is available as a package for R on both Windows and Linux/Unix, and as such may be considered platform-independent.

In this study, DetectiV produced the correct result in 55 out of 56 arrays, by filtering for viruses with a mean normalized log ratio greater than 1 and then sorting by *p* value. We make the distinction here between biological and statistical significance. A statistically significant result may be obtained by a group of oligos that display intensities only marginally larger than the negative control (in this case the global median intensity). This is demonstrated by human herpesvirus 5 on array GSM40820 (Table 3). However, we know that from a biological perspective, we would expect to see intensities far higher than the negative control, and that intensities only marginally higher result from low homology between the probe and the sample. We can therefore use the statistical significance (*p* value) in combination with our idea of biological significance (the mean normalized log ratio) to successfully call the correct result in over 98% of the arrays.

In the majority of cases there is a clear difference in the *p* value, the mean normalized log ratio, or both, between the correct hit and subsequent hits, allowing for both automatic and manual detection of true and false positives. However, this does require careful interpretation. Both DetectiV and E-Predict predict multiple, significant matches on all of the arrays. When using DetectiV, it is only when looking for major changes between the top hit and subsequent hits, in terms of *p* value or mean log ratio, that it is possible to separate the true positives from the false positives. In many cases, using automatic rules will result in the correct result; however, there will inevitably be borderline cases where human inspection of the results is required. This is all the more important when considering the possible economic impacts of a false positive for certain species. At present, the safest way to employ such arrays, and their analysis methods, may be simply as a first step towards identifying infectious agents, informing researchers about which viruses they should test for using more conventional methods.

The results from the application of DetectiV to the SARS dataset are encouraging. Here, oligos designed to SARS were not present on the array. However, using a simple NCBI blast search, it was possible to extend the range of viruses covered by the array to include SARS - 61 existing oligos showing significant homology to the SARS genome. On application of DetectiV to the updated data, SARS was the top hit. Not only does this offer the promise of being able to extend the coverage of the array without adding further oligos, it also suggests that it is possible to detect viruses without having any unique oligos. This may inform the oligo selection process - it may be equally desirable to have multiple, non-unique oligos to represent a species as it is to have a few that are unique.

The results from the application of DetectiV to a second dataset are also encouraging, with the correct result being the top hit in all 12 cases. Of particular interest is the ability of the array, and DetectiV, to distinguish not only between separate viral species, but also between different subtypes of FMDV. It should be noted that in order to apply DetectiV to a second dataset from a completely different array to the first dataset, the user only has to change the GEO accession number and the number of arrays within that dataset. This compares favorably with E-Predict, which would require a separate training dataset from the second array, the calculation of a large and complex sequence similarity matrix and the optimization of several parameters.

There are a number of ways in which DetectiV may be developed. In terms of visualization, better browsing capabilities of the barplots would be desirable, perhaps using a web-interface. In terms of the analysis, we may borrow ideas from gene expression arrays. For example, limma uses an empirical Bayes method to shrink each gene's standard error towards a common value, and has been shown to perform better than standard statistical methods [12]. It may be that we can apply a similar method here to shrink the standard error for each virus species towards a common value, thus increasing sensitivity. It may also be possible to apply multiple-testing procedures to the resulting *p* values. The Bonferroni correction may be appropriate, in which the *p* values are multiplied by the number of comparisons, or a more conservative approach may be needed, such as that suggested by Benjamini and Hochberg [19], in order to control the false discovery rate.

In conclusion, DetectiV is a highly accurate tool for the analysis of pathogen detection microarray data, offering simple but powerful visualization, normalization and significance testing functions. DetectiV performs better than previously published software on a publicly available microarray dataset. DetectiV is available as a package for R, a platform-independent statistical software package, and requires little configuration or customization. It is released under the GNU General Public License and may be downloaded from the DetectiV website [20].

Abbreviations

DFA, direct fluorescent antibody; FMDV, foot-and-mouth disease virus; GEO, Gene Expression Omnibus; IBV, infectious bronchitis virus.

Authors' contributions

Michael Watson wrote and tested the DetectiV software. Juliet Dukes and Abu-Bakr Abu-Median designed the visualization styles in DetectiV, tested the software and produced the data in GSE8746. Donald King and Paul Britton tested the software and helped produce the data in GSE8746.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a table listing the top hit for all 56 arrays using both the Detectiv and E-Predict.dist methods. DetectiV produced a correct result in 55 out of 56 arrays, and E-Predict produced a correct result in 53 out of 56 arrays.

Acknowledgements

This work was supported by the Department of Environment, Food and Rural Affairs (DEFRA) project codes SE4102, SD0443, SE1120 and the Biotechnology and Biological Sciences Research Council (BBSRC). Some of the oligonucleotide probes were provided by Dr M Banks of Veterinary Laboratories Agency (VLA).

References

1. Lapa S, Mikheev M, Shchelkunov S, Mikhailovich V, Sobolev A, Blinov V, Babkin I, Guskov A, Sokunova E, Zasedatelev A, *et al.*: **Species-level identification of orthopoxviruses with an oligonucleotide microchip.** *J Clin Microbiol* 2002, **40:**753-757.
2. Boonham N, Walsh K, Smith P, Madagan K, Graham I, Barker I: **Detection of potato viruses using microarray technology: towards a generic method for plant viral disease diagnosis.** *J Virol Methods* 2003, **108:**181-187.
3. Song Y, Dai E, Wang J, Liu H, Zhai J, Chen C, Du Z, Guo Z, Yang R: **Genotyping of hepatitis B virus HBV by oligonucleotides microarray.** *Mol Cell Probes* 2006, **20:**121-127.
4. Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J, *et al.*: **Viral discovery and sequence recovery using DNA microarrays.** *PLoS Biol* 2003, **1:**E2.
5. Perrin A, Duracher D, Perret M, Cleuziat P, Mandrand B: **A combined oligonucleotide and protein microarray for the codetection of nucleic acids and antibodies associated with human immunodeficiency virus, hepatitis B virus, and hepatitis C virus infections.** *Anal Biochem* 2003, **322:**148-155.
6. Sergeev N, Distler M, Courtney S, Al-Khaldi SF, Volokhov D, Chizhikov V, Rasooly A: **Multipathogen oligonucleotide microarray for environmental and biodefense applications.** *Biosens Bioelectron* 2004, **20:**684-698.
7. Burton JE, Oshota OJ, North E, Hudson MJ, Polyanskaya N, Brehm J, Lloyd G, Silman NJ: **Development of a multi-pathogen oligonucleotide microarray for detection of Bacillus anthracis.** *Mol Cell Probes* 2005, **19:**349-357.
8. Lemarchand K, Masson L, Brousseau R: **Molecular biology and DNA microarray technology for microbial quality monitoring of water.** *Crit Rev Microbiol* 2004, **30:**145-172.
9. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, Wang D, DeRisi JL: **E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns.** *Genome Biol* 2005, **69:**R78.
10. **The R Project for Statistical Computing** [http://www.R-project.org]
11. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **510:**R80.
12. Smyth GK: **Limma: linear models for microarray data.** In *Bioinformatics and Computational Biology Solutions using R and Bioconductor* Edited by: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York: Springer; 2005.
13. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy - analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20:**307-315.
14. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles - database and tools.** *Nucleic Acids Res* 2005, **33:**D562-566.
15. **Full Results from DetectiV for GSE2228** [http://detectiv.sf.net/table3.html]
16. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2000, **28:**10-14.
17. **Full Results from DetectiV for GSE8746** [http://detectiv.sf.net/VA1out.html]
18. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL: **Microarray-based detection and genotyping of viral pathogens.** *Proc Natl Acad Sci USA* 2002, **99:**15687-15692.
19. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Stat Soc B* 1995, **57:**289-300.
20. **The DetectiV Website** [http://detectiv.sf.net/]

BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 1 2015, pages 114–115
doi:10.1093/bioinformatics/btu590

Sequence analysis

Advance Access publication August 29, 2014

poRe: an R package for the visualization and analysis of nanopore sequencing data

Mick Watson[1,*], Marian Thomson[2], Judith Risse[2], Richard Talbot[1], Javier Santoyo-Lopez[2], Karim Gharbi[2] and Mark Blaxter[2]

[1]Edinburgh Genomics, The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Edinburgh EH25 9RG and [2]Edinburgh Genomics, Institute of Evolutionary Biology, Ashworth Laboratories, University of Edinburgh, Edinburgh EH9 3JT, UK

Associate Editor: Inanc Birol

ABSTRACT

Motivation: The Oxford Nanopore MinION device represents a unique sequencing technology. As a mobile sequencing device powered by the USB port of a laptop, the MinION has huge potential applications. To enable these applications, the bioinformatics community will need to design and build a suite of tools specifically for MinION data.

Results: Here we present poRe, a package for R that enables users to manipulate, organize, summarize and visualize MinION nanopore sequencing data. As a package for R, poRe has been tested on Windows, Linux and MacOSX. Crucially, the Windows version allows users to analyse MinION data on the Windows laptop attached to the device.

Availability and implementation: poRe is released as a package for R at http://sourceforge.net/projects/rpore/. A tutorial and further information are available at https://sourceforge.net/p/rpore/wiki/Home/

Contact: mick.watson@roslin.ed.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 4, 2014; revised on August 25, 2014; accepted on August 26, 2014

1 INTRODUCTION

Relative to first- and second-generation sequencing technologies, single-molecule sequencing is a new science, with only Helicos (Bowers *et al.*, 2009), Pacific Biosciences (Eid *et al.*, 2009) and Oxford Nanopore (ONT) being widely available. Even within the field of single-molecule sequencing, ONT's nanopore sequencing technology represents a new paradigm; while both Helicos' and Pacific Biosciences' sequencing technologies measure incorporation events into a second strand, ONT's MinION and GridION systems measure a single molecule of DNA as it passes through a protein nanopore. In addition, the MinION is the world's first mobile DNA sequencer; it is powered by a laptop's USB port and measures ~4 inches in length. Recently, ONT opened up the MinION access programme, enabling researchers to use the device for the first time.

The ultra-low-cost and mobile nature of the MinION device opens up a huge number of applications. However, users of the device are faced with a number of informatics challenges. Users of the MinION must buy a high-specification Windows laptop, and thus there is a need for Windows-based software to handle the data. The MinION outputs binary files in the HDF5 format (http://www.hdfgroup.org/HDF5/). These contain raw data from the sequencer, which are then processed by a cloud-based base caller called 'metrichor'. The subsequent called sequence files are also in HDF5 format (with the extension .FAST5). It is not uncommon for users to be presented with 30–50 000 HDF5 files (.fast5), with no software with which to access the data. Furthermore, data from all runs are stored in a single directory, with no subdirectories, and users find themselves needing to manipulate thousands of files manually, which takes time and is error prone.

We have developed poRe, a package for the statistical package R (http://www.R-project.org/; R Core Team, 2014), which enables users to manipulate MinION FAST5 files into run folders, extract FASTQ, gather statistics on each run and plot a number of key graphs, such as read-length histograms and yield-over-time. Crucially, as a package for R, poRe is cross-platform and has been tested on Windows, Linux and MacOSX. The Windows version enables users to run poRe on the MinION laptop itself, rather than copying the data to a Linux server to process with Perl or Python. This key feature brings users closer to true mobile DNA sequencing.

2 METHODS

2.1 Data format

The FAST5 HDF5 files contain a number of hierarchical groups, datasets and attributes, and these are described in more detail in the Supplementary Information.

2.2 Organization and run statistics

The first task users face is to organize a single MinION folder, which may contain reads from many different runs. We provide the function *copy.runs( )* to help with this. The function reads all FAST5 files within a user-defined directory and extracts both the unique run identifier ('run_id') and the name and version of the base caller. Each read is then copied to a user-defined destination folder, under subfolders defined by the run_id and the name and version of the analysis. The latter is key, as each raw read may be base called many times by different versions of the metrichor base caller.

Embedded within each FAST5 file are a number of key statistics about the reads. These can be extracted for all reads in a run by the function *read.fast5.info( )*. This returns a data frame with 24 columns of metadata for each read. The function *run.summary.stats( )* can be used to extract

*To whom correspondence should be addressed.

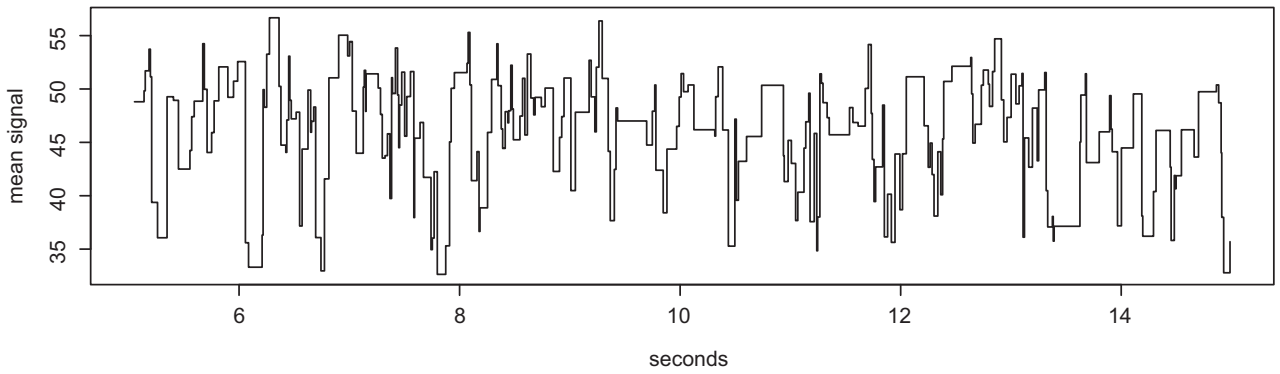


Fig. 1. An example output from the *plot.squiggle* function. Plotted are the events data extracted from a single read. The *y*-axis is the mean electronic signal reported for the pore, and the *x*-axis is time in seconds

key summary statistics, such as maximum, minimum and mean read lengths.

2.3 FASTQ and FASTA extraction

Once the data are organized, users may wish to extract FASTQ/A data. This can be done using the *extract.run.fastq( )* and *extract.run.fasta( )* functions. For each FAST5 read in a given directory, this function will extract the template, complement and 2D FASTQ/A data where they exist and write these to individual FASTQ/A files.

2.4 Data exploration

We provide a number of functions that allow users to explore the data visually. Histograms of read length can be created using the *plot.length.histogram( )* function. This plots histograms for the template, complement and 2D read lengths (Supplementary Fig. S1).

The *plot.cumulative.yield( )* function can be used to plot cumulative yield of the run over time, and sums up the template, complement and 2D read lengths over time in seconds since the analysis began (Supplementary Fig. S2).

Finally, the MinION device consists of a number of channels, each of which should contain a single nanopore. Users can count and plot the number of reads per channel for a run, using *plot.channel.reads( )*, and sum and plot the yield per channel, using *plot.channel.yield( )*. Both of these can be potentially used to diagnose problems in particular areas of the flowcell.

2.5 Extracting and plotting events

The raw data from the MinION are the information about the electronic signal measured as each single molecule of DNA passes through the protein nanopore. It is these data that are converted to sequence data by the metrichor agent. However, the raw events data are also available and can be extracted using the function *get.events( )*. This will extract the thousands of events for both the template and complement for a particular read. The events data may then be visualized using the *plot.squiggle( )* function (see Fig. 1).

3 DISCUSSION

We have written poRe, an R package that enables users to more easily manipulate, summarize and visualize MinION nanopore sequencing data. As a package for R, poRe is available for both Windows and Linux, and crucially the Windows version will allow data analysis on the mandatory Windows laptop on which the MinION depends. In addition, R is now a popular statistical package among biologists, who may feel comfortable using poRe through the R user interface.

poRe is one of the first bioinformatics packages to offer this necessary functionality. poretools (Loman and Quinlan, 2014), a toolkit written in Python, offers similar functionality, although each software has a different set of (overlapping) functions. A table comparing feature sets is available in the Supplementary Information. The cross-platform nature of poRe, its ease of installation and poRe's ability to organize folders of FAST5 files make poRe an important tool for users of the MinION device.

ACKNOWLEDGEMENT

The authors would like to thank Oxford Nanopore for granting Edinburgh Genomics access to the MinION Access Programme (MAP).

Funding: Edinburgh Genomics is partly supported through core grants from the National Environmental Research Council (NERC R8/H10/56), Medical Research Council (MRC MR/K001744/1) and The Biotechnology and Biological Sciences Research Council (BBSRC BB/J004243/1).

Conflict of interest: none declared.

REFERENCES

R Core Team. (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Bowers,J. *et al.* (2009) Virtual terminator nucleotides for next-generation DNA sequencing. *Nat. Methods*, **6**, 593–595.

Eid,J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Loman,N.J. and Quinlan,A.R. (2014) Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, pii, btu555.

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 6 2009, pages 832–833
doi:10.1093/bioinformatics/btp059

Systems biology

CORNA: testing gene lists for regulation by microRNAs

X. Wu* and M. Watson

Bioinformatics Group, Institute for Animal Health, Compton, RG20 7NN, UK

Received on November 19, 2008; revised on January 06, 2009; accepted on January 25, 2009

Advance Access publication January 29, 2009

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: With the increasing use of post-genomics techniques to examine a wide variety of biological systems in laboratories throughout the world, scientists are often presented with lists of genes that they must make sense of. A consistently challenging problem is that of defining co-regulated genes within those gene lists. In recent years, microRNAs have emerged as a mechanism for regulating several cellular processes. In this article, we report on how gene lists and microRNA targets data may be integrated to test for significant associations between gene lists and microRNAs.

Results: We discuss CORNA, a package written in R and released under the GNU GPL, which allows users to test gene lists for significant microRNA–target associations using one of three separate statistical tests, to link microRNA targets to functional annotation and to visualize quantitative data associated with those data.

Availability: CORNA is available as an R package from http://corna.sf.net

Contact: xikun.wu@bbsrc.ac.uk

1 INTRODUCTION

Experiments involving post-genomics technologies such as microarrays, proteomics and systems biology often present scientists with gene lists that they must attempt to make sense of. Several software packages exist that allow scientists to assign functional annotation to gene lists, and to assign statistical significance to those associations. These include tools for associating genes with biological ontologies (e.g. Falcon and Gentleman, 2007) and with biological pathways (e.g. Salomonis *et al*., 2007).

A particular challenge is that of assessing which genes in a given gene list are co-regulated. miRBase (Griffiths-Jones *et al*., 2006), a database of all known microRNAs, has been created and there have been several published software tools that attempt to predict the targets of microRNAs (Brennecke *et al*., 2005). An excel-based tool (Creighton *et al*., 2008) has been produced for linking microarray data to microRNA targets information.

Here we describe CORNA, a package for R that allows scientists to analyse gene lists in the context of microRNA–target predictions. Methods exist to test for significant microRNA–target relationships in gene lists, and to test for significant associations between microRNAs and pathways and GO terms. The software is flexible and can read data from public databases or from a scientists own data files. CORNA is released as open-source under the GNU GPL.

*To whom correspondence should be addressed.

2 FLOW OF INFORMATION

Central to the flow of information through CORNA is the gene list from which the user may test for significant microRNA–target associations. The user may also start with a microRNA, find genes that are associated with that microRNA and then test that gene list for significant associations with KEGG pathways or GO terms. The user may also plot quantitative data associated with the targets of a particular microRNA.

2.1 Inputs

CORNA exclusively uses R vectors and data frames. CORNA includes functions for reading microRNA–target data directly from miRBase and microRNA.org (Betel *et al*., 2008). There are also helper functions to read gene and GO term data using biomaRt (Durinck *et al*., 2005); microarray data directly from GEO (Barrett *et al*., 2008); and pathway data directly from KEGG (Kanehisa *et al*., 2004).

2.2 Methods

CORNA employs three complementary statistical methods for enrichments analysis of relationships within lists of genes. These are the HyperGeometric test, Fisher's exact test and the χ^2-test.

2.3 Outputs

If the user tests a gene list for significant microRNA associations, then the output is an R data frame with one row per microRNA, the observed and expected frequencies from sample and population, and the range of user-selected *P*-values.

Where the user begins with a particular microRNA, the targets information is used to create a gene list and that gene list is tested for enrichment of pathways and GO terms.

There is also a range of plotting functions for plotting quantitative data associated with microRNA targets.

3 EXAMPLE ANALYSIS

3.1 Using CORNA to test for enrichment of microRNA–target relationships in a gene list

The list in this example, *tsam*, consists of 1000 ensembl transcript ids; 940 of these were chosen at random, then 30 predicted targets for two microRNAs were added. The example assumes that the file 'arch.v5.txt.mus_musculus.zip' has been downloaded from miRBase targets.

```
targets <- miRBase2df.fun(
             file="arch.v5.txt.mus_musculus.zip")
             data(CORNA.DATA)
res <- corna.test.fun(
          x=tsam,
          y=unique(targets$tran),
          z=targets,
          p.adjust="BH")
```

The only two microRNAs with a significant adjusted *P*-values are those used to bias the transcript list. The user may work with genes simply by converting the transcript list to microRNA–gene relationships using the *BioMart2df.fun* and *corna.map.fun* functions.

3.2 Using CORNA to test for KEGG pathways associated with a microRNA list

The microRNA used in this example is 'mmu-mir-155', and we use the predicted targets from miRBase to test for enrichment of KEGG pathways.

```
tran2gene <- BioMart2df.fun(
               biomart="ensembl",
               dataset="mmusculus_gene_ensembl",
               col.old=c("ensembl_transcript_id",
                         "ensembl_gene_id"),
               col.new=c("tran", "gene"))
mir2gene <- corna.map.fun(targets, tran2gene,
              "gene",
              "mir")
gvec <- corna.map.fun(mir2gene,
                      "mmu-mir-155",
                      "mir",
                      "gene")
gene2path <- KEGG2df.fun(org="mmu")
gvec <- intersect(gvec, unique(gene2path$gene))
test <- corna.test.fun(
            gvec,
            unique(gene2path$gene),
            gene2path,
            hypergeometric=T,
            fisher=T,
            fisher.alternative="greater",
            min.pop=10,
            sort="fisher")
```

We first convert the microRNA–transcript relationship to a microRNA–gene relationship using the *BioMart2df.fun* and *corna.map.fun* functions. We then find those genes predicted to be targets of mmu-mir-155. The next stage is to use the *KEGG2df.fun* function to obtain links between genes and pathways from KEGG for *Mus musculus*. Finally, we set the sample to be only those genes targeted by mmu-mir-155 that have a pathway link, and perform hypergeometric and Fisher's exact tests for the KEGG pathways involved. The top five pathways can be seen in Table 1.

Table 1. Top five significant pathways from CORNA for mmu-mir-155

ID	Description	Expected	Observation	*P*-value
00190	Oxidative phosphorylation	5	12	0.002
00400	Phenylalanine etc. biosynthesis	0	3	0.003
00500	Starch and sucrose metabolism	2	6	0.012
05020	Parkinson's disease	5	10	0.016
04010	MAPK signaling pathway	9	15	0.044

4 SUMMARY

With increasing use of large-scale post-genomics techniques, scientists are often presented with lists of genes. MicroRNAs have emerged as an important regulator of gene function. In this article, we have shown that CORNA can be used to test for significant associations between genes, microRNAs, pathways and GO terms. CORNA can also be used to plot quantitative data associated with microRNA targets. CORNA is flexible and can read data from public databases or from a user's own files. CORNA has been tested on both Microsoft Windows and Red Hat Linux. CORNA is released under the GNU GPL and is available from http://corna.sf.net.

Funding: BBSRC (core strategic grant of the Institute for Animal Health).

Conflict of Interest: none declared.

REFERENCES

Barrett,T. *et al*. (2008) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D5–D15.

Betel,D. *et al*. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–D153.

Brennecke,J. *et al*. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.

Creighton,C.J. *et al*. (2008) A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *RNA*, **14**, 2290–2296.

Durinck,S. *et al*. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.

Griffiths-Jones,S. *et al*. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.

Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Salomonis,N. *et al*. (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, **8**, 217.

Chapter 20

Large-Scale Integration of MicroRNA and Gene Expression Data for Identification of Enriched MicroRNA–mRNA Associations in Biological Systems

Preethi H. Gunaratne, Chad J. Creighton, Michael Watson, and Jayantha B. Tennakoon

Abstract

The discovery of microRNAs (miRNAs) revealed a hidden layer of gene regulation that is able to integrate multiple genes into biologically meaningful networks. A number of computational prediction programs have been developed to identify putative miRNA targets. Collectively, the miRNAs that have been discovered so far have the potential to target over 60% of genes in our genome. A minimum of six consecutive nucleotides in the 5′-seed (nucleotides 2–8) in the miRNA must bind through complimentary base pairing to the 3′-untranslated (3′-UTRs) of target genes. Given the small sequence match required, a given miRNA has the potential to target hundreds of genes and a given mRNA can have 0–50 miRNA binding sites. The low-throughput nature of the query design (gene by gene or miRNA by miRNA) and a fairly high rate of false positives and negatives uncovered by the limited number of functional studies remain as the major limitations. Programs that integrate genome-wide gene and miRNA expression data determined by microarray and/or next-generation sequencing (NGS) technologies with the publicly available target prediction algorithms are extremely valuable on two fronts. First, they allow the investigator to fully capitalize on all the data generated to reveal new genes and pathways underlying the biological process under study. Second, these programs allow the investigator to lift a small network of genes they are currently following into a larger network through the integrative properties of miRNAs. In this chapter, we discuss the latest methodologies for determining genome-wide miRNA and gene expression changes and three programs (Sigterms, CORNA, and MMIA) that allow the investigator to generate short lists of enriched miRNA:target mRNA candidates for large-scale miRNA:target mRNA validation. These efforts are essential for determining false positive and negative rates of existing algorithms and refining our knowledge on the rules of miRNA–mRNA relationships.

1. Introduction

MicroRNAs (miRNAs) are small ~22 nucleotide noncoding RNAs that have been predicted to target >60% of the genes in our genome to mediate posttranscription gene silencing (1, 2). The

Silvia Monticelli (ed.), *MicroRNAs and the Immune System: Methods and Protocols*, Methods in Molecular Biology, vol. 667, DOI 10.1007/978-1-60761-811-9_20,

key determinants for miRNA–mRNA target associations lie in the 5′-seed region (nucleotides 2–8) in miRNA and the 3′-untranslated region (3′-UTR) of mRNA targets (2). The miRNA–mRNA target association is catalyzed mainly by the action of Argonaute (Ago) family of proteins in the RNA-induced silencing complex (RISC) (3). Base pairing of at least six consecutive nucleotides within the 5′-seed of the miRNA with the target site on the mRNA is reported to be required at a minimum. However, binding can occur through the entire length of the miRNA. miRNA–mRNA duplexes that form with perfect or near perfect complementarity have been shown to result in mRNA cleavage between nucleotides 10 and 11 (4) of the miRNA resulting ultimately in mRNA cleavage and decay (4, 5). By contrast, when binding occurs through imperfect complementarity, the mRNA target is generally kept intact and silencing occurs through translational repression (6).

With the advent of microarray and next-generation sequencing (NGS) technologies in the postgenome era, it is now possible to determine genome-wide miRNA–mRNA associations that are significant to specific cellular contexts or systems such as the immune system. A number of target prediction algorithms, which are primarily based on searches for matches between miRNA seed sequences and 3′-UTRs of genes, have been developed and freely available (7). Such programs offer users the possibility of quickly searching for potential targets on a miRNA by miRNA basis or potential miRNAs on a gene-by-gene basis. However, these approaches are too cumbersome and do not offer optimal solutions to integrate the glut of microarray (gene and miRNA expression) and sequencing (mRNA-seq and miRNA-seq) data that is becoming available on a daily basis. More recently, several groups have written programs and software packages to address this issue and offer solutions for the large-scale three-way integration of gene expression data, miRNA expression data and miRNA–mRNA target predictions (8). These programs offer the users the possibility of reaping the full benefit of these genome-wide studies. It is becoming increasingly clear that miRNAs are very different from the traditional transcriptional repressors that we are familiar with. Overexpression and loss-of-function studies suggest that most miRNAs have only a limited influence on their target genes (approximately two- to ten-fold repression) on its own. It appears that the main role of miRNAs is to fine-tune gene expression by coordinately downregulating multiple genes within and across pathways to integrate them into meaningful networks in relation to specific cellular states. The question then is what is the impact of global shifts in miRNA profiles on the transcriptome and proteome of a given cellular state. Furthermore, when aiming to assess the role of a given miRNA in relation to a specific biological process,

it is essential to consider its impact on all of its targets. Consequently, programs that integrate expression data with target prediction data are vital to understand the role of miRNAs in the immune system. In this chapter, we examine in detail three programs that allow the large-scale integration of genome-wide expression and miRNA target prediction data.

2. Materials

2.1. Gene Expression Data

2.1.1. Microarrays

While classical northern blotting and quantitative real-time PCR continue as techniques used for gene expression studies of single or small sets of genes over the past two decades, high-throughput microarray-based techniques have been increasingly applied in this field to measure several thousands of genes at a time. Microarray technology was first described by Schena et al. in 1995 (9). Over the years, DNA chip based technologies have widely demonstrated the power of this high-throughput parallel synthesis based method. Microarray DNA chips contain thousands of probes arranged on a regular pattern. Microarrays produce quantitative gene expression data based on relative dye intensities corresponding to DNA hybridized to probes immobilized on chips (10). A typical microarray-based experiment consists of preparing a DNA chip based on target DNAs, generating a hybridization solution containing a mixture of fluorescently labeled cDNAs, incubating fluorescently labeled cDNAs with DNA chip followed by data detection based on laser technologies, and finally computer assisted statistical testing and data analysis. To disseminate data analyzed by researchers for public use, microarray data can be stored in NCBI microarray data repository Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) (11) using a standardized framework, termed microarray markup language (MAML).

MAML employs a standard format to describe microarray experiment details, which include experimental design, array design, samples, hybridization procedures and parameters, images, quantitation, and controls.

Several commercial producers have introduced microarrays with different features. The microarrays available in the current market differ from one another in terms of the technologies utilized for fabrication and their probe design architecture. Some of the popularly known commercial manufacturers are given below:

Affymetrix GeneChip (http://www.affymetrix.com). Affymetrix was one of the first microarrays to appear in the market (12). Unlike in the case of traditional microarrays where cloning libraries are used for probe design, Affymetrix employs an in silico light directed synthesizing technology to produce probes on a glass chip (10).

Bypassing management of clone libraries and ability to synthesize highly ordered DNA oligomers in silico are the two distinct advantages of the Affymetrix design.

Agilent (http://www.agilent.com) employs an inkjet-based method to print whole cDNA or oligos on chips (13). Chips produced by Agilent offer 60-mer probes compared to the 25-mer probes offered by Affymetrix (10).

Nimblegen (http://www.nimblegen.com/) (14) uses a digital light processor to synthesize microarrays, apart from their use in transcriptome analysis. NimbleGen chips containing specific sequences are used to capture large genomic fragments, which can be subject to further analysis using Nimblegens GS FLX sequencing system (10).

CombiMatrix (http://www.combimatrix.com) offers custom arrays generated by a powerful computer-directed semiconductor microelectrode based on chip synthesis method, which can be programmed to generate a given array of oligonucleotides on chips (15). Signal detection can be carried out by either laser scanning or electrochemical methods (10).

Illumina bead array (http://www.illumina.com) (16) conventional microarrays are manufactured by spotting oligonucleotides on two-dimensional substrates (17). On the contrary, Illumina bead based arrays are produced by means of random assembly of bead pools on a patterned substrate (17). Illumina's technology offers higher oligo densities on their chips and thus higher throughputs by virtue of the intrinsic size of the beads and patterned substrates compared to conventional chips.

While array-based technologies and applications continue to grow, a plethora of information would be available for researchers through GEO in future. This would be a very valuable tool to facilitate cross-reference samples, identify signatures associated with disease, personalize medicine, and most importantly provide a global view of all biological processes through a platform for systematic in depth analysis of DNA and RNA variation.

2.2. miRNA Expression Data

2.2.1. MicroRNA Microarrays

The overall approach of miRNA profiling through microarrays remains similar to the approach employed in microarrays for gene expression profiling. Mature miRNAs are isolated and purified from tissue or cell samples using classical Trizol-based isolation or commercially available kits. The purified fragment of RNA is enriched and labeled. Array probes are designed by using locked nucleic acid (LNA) or chemically modified oligos and spotted on microarrays. Hybridization is then carried out and signal intensities measured using a laser scanner. Finally, quantification and data analysis is carried out using computer software. Unlike in the case of mRNA arrays designing arrays for miRNAs is challenging in that arrays must be designed to discriminate between the mature miRNAs and their precursors, miRNA microarrays should

be capable of detecting subtle differences of even a single-base difference of mature sequences (18). Short sequence length of 18–25 nt of mature sequences and wide range of melting temperatures (T_m) of mature miRNAs are significant problems in miRNA microarray design (19). In spite of the challenges, several miRNA microarrays have been designed and are currently available commercially. Synthetic oligonucleotides or cDNA fragments are used in miRNA microarray probe design. More recently, synthetic oligonucleotides with chemical modifications providing high molecular affinities facilitating hybridization have been employed. AT-rich probes are known to show lesser hybridization affinity compared to GC-rich probes (20). Higher degrees of sensitivity can be achieved by introduction of A/T analogs, which enhance overall duplex stability (21). Substitution of A and T with 2^0-O-methyl-2,6-diaminopurineand2^0-O-methyl-5-methyluridine, respectively, has shown two- to threefold increases in relative hybridization (22). LNAs first described by Wengel and coworkers in 1998 are a novel class of conformationally restricted oligonucleotide analogs, which show high thermal stabilities toward complimentary RNA and DNA (23). Chemically engineered LNAs have nucleotide analogs containing a bridging methylene group between C4′ and O2′ of the ribose ring (24). High thermal stabilities of LNAs bound to complimentary nucleic acid facilitates the design of short probes with excellent mismatch discrimination. Some of the commercially produced miRNA microarrays are discussed next.

2.2.1.1. Agilent miRNA Microarray (http://www.home.agilent.com)

Agilent miRNA microarrays are produced using unique chemically unmodified probes. Chemically unmodified oligos are immobilized on an array platform by means of a short stilt, and to the 5′ end of the anchored oligo a G residue is included and an extended hairpin attached, the 3′ end of the sample miRNAs are labeled by means of a Cy molecule attached to a C residue. When sample is introduced, hybridization takes place and the 5′ G residue of the probe complimentary to the 3′ Cy labeled C residue binds resulting fluorescence. The hairpin functions as a bridge connecting the 5′ end of the anchored oligo and the 3′ end of the hybridized miRNA. Agilent claims that the inclusion of the G residue to 5′ end of the probe increases stability of binding to target miRNA and the hairpin destabilizes probe hybridization to larger nontarget RNAs and hence provides a higher degree of specificity. Agilent's G44071A human miRNA microarray platform uses sequences from Sanger miRNA database (miRBase) version 12 and is capable of detecting unique 866 human and 89 viral miRNAs. Agilent also produces several arrays in the G44 series for human mouse and rat miRNAs, which use different versions of the Sanger database ranging from version 9.1 to 12.0 as the reference source for sequences. In Agilent miRNA

arrays, 40–60 mer unmodified oligonucleotides are directly synthesized on the array by Agilent's proprietary SurePrint inkjet technology. A unique feature of Agilent's technology is the use of end labeling instead of conventional polymerase based methods where sample nucleotide damage within the substrate has been an issue. End labeling is insensitive to nucleotide damage and is particularly advantageous when testing preserved or chemically treated samples. Agilent's platform requires only small input amounts in the 100 ng range of total RNA due to the high-yield end labeling method. As the labeling method does not require size fractionation or amplification, undesired bias introduced from these two steps is eliminated (25).

2.2.1.2. Exiqon LNA Microarrays (http://www.exiqon.com)

Exiqon uses melting temperature (T_m) matched LNA probes in their miRNA microarray design. Exiqon's miRNA microarrays are marketed under the name miRCURY LNA™. In addition to probes for miRBase sequences, which the Exiqon system uses as a reference for their microarrays, Exiqon arrays contain probes called mirPLUS™ capture probes, which target proprietary miRNAs that have been defined by Exiqon company through cloning and sequencing of human normal and diseased tissues. Through these proprietary sequence probes, scientists would be able to gain unique information about miRNAs, which have not been defined elsewhere. As of August 2009 in the Exiqon Web site, a typical miRCURY LNA™ was listed as being capable of capturing 854 mature human miRNAs, 80 mature viral miRNAs, and 428 mature Exiqon-defined human mirPLUS™ miRNAs (26).

2.2.1.3. Invitrogen (http://www.invitrogen.com)

Invitrogen offers the NCode™ Human miRNA Microarray Kit V3 and NCode™ Multi-Species miRNA Microarray Kit V2 as integrated miRNA profiling systems, which include reagents for RNA isolation labeling and array hybridization. As of the date of writing this chapter (30 August 2009), it was listed in the Invitrogen Web site that the Human miRNA Microarray Kit V3 contains probe sequences targeting nearly all of the known human miRNAs in the Sanger miRBase as well as probe sequences for 373 novel putative miRNAs. The Multi-Species version was listed as having probes for the Sanger miRBase Sequence Database, Release 9.0, for human, mouse, rat, *Drosophila melanogaster*, *Caenorhabditis elegans*, and Zebrafish. Each NCode™ microarray slide comes fully blocked and ready to use. In case where starting material has concentrations <500 ng total RNA or equivalent cells/tissue, Invitrogen provides a miRNA amplification kit called the NCode™ miRNA amplification system. Once the total RNA is extracted and ready for hybridization, labeling can be carried out using an NCode™. Rapid miRNA labeling system, which is

based on a poly-A tailing reaction on RNA molecules followed by ligation of dye labeled Alexa Flour™ DNA polymer by means of an oligoDT bridge. Invitrogen offers the choices of employing either preprinted or self-printed microarrays using NCode™ Human or multispecies microarray probes for experiments. Analysis of results can be performed by means of NCode™ profiler software. Invitrogen also provides a range of reagents under the NCode™ name for verification and further analysis of results by qPCR (27).

2.2.1.4. LC Sciences (http://www.lcsciences.com)

With the LC Sciences μParaFlo microfluidic miRNA Microarray chips assays can be performed with a minimum of 5 μg total RNA (28). The mirVana Isolation Kit (Ambion) is recommended. The small RNA (<300 nt) fraction is size fractionated with YM-100 Microcon Centrifugal Filter Device (Millipore) and 3′-extended with a poly-A tail by poly-A polymerase. An oligonucleotide tag is ligated to the poly-A tail for subsequent fluorescent dye staining. This platform allows dual labeling using Cy5 and Cy3 tags to label two RNA samples to be compared in dual-sample experiments. Hybridization is performed overnight on a μParaFlo microfluidic chip using a microcirculation pump (Atactic Technologies). On the microfluidic chip, each detection probe consists of a chemically modified nucleotide "coding" segment complementary to target miRNA (from miRBase, http://microrna.sanger.ac.uk/sequences/) or other RNA (control or customer defined sequences) and a spacer segment of polyethylene glycol to extend the "coding" segment away from the substrate surface. The detection probes are synthesized in situ with photogenerated reagent (PGR) chemistry on a Digital Light Projector (Texas Instruments) based synthesis system (29). Flexible DNA chip synthesis is gated by deprotection using solution photogenerated acids. The hybridization melting temperatures are balanced by adjusting length and chemical modifications of the detection probes (29). Hybridization is carried out in 100 μL 6× SSPE buffer (0.90 M NaCl, 60 mM Na_2HPO_4, 6 mM EDTA, pH 6.8) containing 25% formamide at 34°C followed by a stringent wash at 52°C. Hybridization images are collected with a laser scanner (GenePix 4000B, Molecular Device) and signal intensity values extracted using ArrayPro image processing software (MediaCybernetics). Data analysis is carried out by first subtracting the background and then normalizing with a cyclic LOWESS filter (locally weighted regression). For two-color experiments, the ratio of two sets of detected signals ($\log_2$ transformed and balanced) and p-values of the t-test are calculated; a p-value of less than 0.01 is used to select significantly differentially detected signal. Data classification is accomplished

by hierarchical clustering based on average linkage and Euclidean distance metric, and visualized with TIGR's Multiple Experimental Viewer (MeV) (30).

Quantile normalization on the channel values is used to normalize two-color data within each chip to make single channel values within and between arrays more comparable and to improve the multiarray data analysis. The single channel normalized values are used in subsequent data analysis. Construction of a dendrogram on the single channel values, both before and after normalization, is recommended to examine the effect of normalization on the treatment differences.

2.2.2. mRNA-seq: Next-Generation Sequencing

Completion of the human reference genome by the international human genome sequencing consortium and US-based Celera genomics was a cornerstone of human scientific endeavor. This achievement clearly paved way for a new exciting era of scientific research. The human genome sequencing project commenced in the year 1990; by 2000 a draft version of the human genome was made available and a completed version was released in the year 2003. During the human genome sequencing project era, the two widely used technologies were the original enzymatic dideoxy sequencing method pioneered by Fred Sanger and colleagues (31) and the Maxam and Gilbert method, which was described during the same year (32). The chemical degradation based Maxam and Gilbert method was particularly used in cases that were not easily resolved by the popular Sanger technique (33). As the human genome project progressed, the need for fast automated sequencers became imminent and companies with commercial interests were quick to step in to make improvements to the Sanger-based technique. In spite of the advances made in the Sanger technique through introduction of automated capillary sequencers, particularly the sample preparation steps, which involved cloning of sequences into bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs) and artifacts related to sample preparation remained obstacles of making Sanger-based sequencing a completely automatable high-throughput method. In view of this fact, several companies came up with novel sequencing technologies, which had massively parallel high-throughput capabilities enabling genome-scale analysis in a relatively short period of time. These sequencing technologies are termed NGS technologies. As of today, three platforms, namely Roche Applied Science 454 platform, the Illumina platform, and Applied Biosystems ABI SOLiD system are widely used in research laboratories. More recently, the Helicos single-molecule sequencing device, HeliScope was released to the market. A brief description of the 454, Illumina and SOLiD systems are given in the following paragraphs.

2.2.2.1. The 454 GenomeSequencer FLX Instrument (Roche Applied Science) (http://www.454.com/)

The 454 FLX pyrosequencer, which was released in 2004, was the first to be introduced to the market as an NGS (34). In pyrosequencing, each time a nucleotide gets incorporated to the nucleotide chain through a polymerizing reaction, pyrophosphate is released, and the released pyrophosphate leads to a series of downstream events, which results in the production of firefly luciferase (35). In the 454 system, DNA fragments are ligated with special adapters. One of the adapters facilitates binding of the DNA molecule to a bead. Beads containing single DNA fragments are subject to emulsion PCR and followed by a denaturation step. Initial amplification of sample DNA is necessary to generate sufficient signal strength in the sequence by synthesis step, which is subsequently carried out on beads containing copies of a given fragment immobilized on an optical fiber chip. In the 454 setup, each bead with its amplified fragment is individually addressable by a CCD camera at the fluorescence detection stage. In the sequence by synthesis stage, polymerase enzyme, primers, and a given labeled nucleotide of known identity are provided to each bead at a time, and the resultant fluorescence due to the pyrosequencing reaction is measured via the optical fibers equipped to a smart camera. By introducing labeled nucleotides of a given kind at each subsequent cycle of the polymerizing reaction, the nucleotides being incorporated to the growing fragment in each cycle can be detected by fluorescence measurement, and the sequence of each fragment can be decoded and assembled using sophisticated computer software. The 454 system is capable of detecting sequences in the 400–500 bp range and generates around 100 MB of data in a single run. A newer improved version of the 454 FLX called Titanium would provide a data output of around 500 MB. High costs of operation and generally low reading accuracy in homopolar stretches have been cited as drawbacks of the 454 system (33).

2.2.2.2. The Illumina (Solexa) Genome Analyzer (http://www.illumina.com/)

The Solexa sequencers were first introduced to the market in the year 2006 (36) and Illumina acquired Solexa in the year 2007 (33). The Solexa system is based on sequencing by synthesis method, which uses a technology called “Reversible termination”. The basic workflow of the Illumina platform involves five main stages. The initial step involves randomly fragmenting DNA and ligating adaptors to random fragments. The second step involves attaching DNA to a special glass slide and is followed by a third step, where solid-phase bridge amplification is carried out using unlabelled nucleotides. The fourth step involves denaturing amplified double-stranded DNA on the slide, and finally the fifth step involves carrying out a PCR using labeled nucleotides and photographing.

Unlike in the case of the 454 instrument where a single variety of nucleotide is incorporated in each cycle of the fluorescence

generating polymerizing step, the Illumina instrument introduces all four labeled nucleotides to the polymerizing reaction at once. However, due to a chemical modification of the nucleotides, each time a nucleotide gets incorporated into the growing DNA chain termination of polymerization occurs. At this stage, a smart CCD camera photographs fluorescence signals resulting from nucleotides, which got incorporated to each individually addressable amplified cluster of DNA fragments, which are generated at the bridge amplification stage. Once photographing of all clusters is completed, termination is reversed and another set of nucleotides are introduced, and once incorporation takes place, the reaction is terminated and clusters are photographed. Eventually all photographic data are analyzed and the sequences are assembled using computer software. The sequence read length achieved by this technology is around 35 bp, and an advantage of this system is its ability to generate huge amounts of data in a single run. The Illumina GA2 sequencers released in 2008 had the ability to generate around 1.5 GB of data in a single read setup and around 3.0 GB of data using a paired run. The ability of the instrument to generate massive amounts of data having short sequence lengths has made this instrument particularly well suited for small RNA based research, which generally does not demand long sequence reads. With various modifications in sample preparation and the use of different reagents, the Illumina platform can be used in a versatile fashion for ChipSeq and Bisulfite sequencing experiments as well.

2.2.2.3. The Applied Biosystems ABI SOLiD System (http://www3.appliedbiosystems.com/)

In contrast to the polymerase reactions used in 454 and Illumina methods, the Applied Biosystems SOLiD technology uses a ligation-based reaction to incorporate fluorescent-labeled nucleotides in the sequencing step (37). However, the Solid system shares similarities with 454 and Illumina as it utilizes an adapter ligated library and emulsion PCR on magnetic beads at the sample preparation stages. The overall work flow of the solid system can be summarized as follows. Initially, an emulsion PCR step is carried out on adapter ligated DNA fragments anchored to magnetic beads to provide sufficient fluorescence intensities during the detection step. The magnetic beads containing the amplified fragments are then transferred to a flow cell slide where a ligation reaction is carried out. The ligation reaction uses a primer, which attaches to the 5 prime end of the adaptor that immobilizes DNA fragments on the magnetic bead. DNA ligase and specific 8 mers whose fourth and fifth bases are specifically encoded with attached fluorescent labels are introduced to the reaction. Fluorescent detection is followed after each extending ligation step. After ligation and detection, a regeneration step in which the 8 mers including the fluorescent labels are removed is carried out and a primer corresponding to a single base displacement (n–1) from

the 3′ end of the adapter attaching the DNA fragment is introduced, and the ligation cycle is followed while the two encoded bases are read. Similar cycles are carried out starting with primers, which correspond to *n*–2, *n*–3, *n*–4, and *n*–5. In each cycle, the encoded two bases are interrogated and data stored. Finally, when all rounds of ligation have been completed, a computer builds the sequence by decoding the stored data as two base pair calls. A distinct advantage of this system is the use of two base pair encoding. As a result of two base pair encoding, it is possible to discriminate between base calling errors, true polymorphisms and single base deletions of the sequence by alignment against a high quality reference. The sequence length in the solid systems is defined in between 25 and 35 by the user. A sequencing run in a SOLiD system can yield 2–4 GB of DNA sequence data (35).

Information regarding the HeliScope instrument is available at http://www.helicosbio.com/ (38). There are also several other companies, which are in the process of manufacturing single-molecule based powerful sequencers employing *state-of-the-art* technologies. The following links provide information regarding these systems, which are either in the developmental phase or are ready to step into the market: VisiGen Biotechnologies (http://visigenbio.com/) (39), Pacific Biosciences (http://www.pacificbiosciences.com/index.php) (40), Sequenom (http://www.sequenom.com) (41), Oxford Nanopore Technologies, UK (http://www.nanoporetech.com/) (42), BioNanomatrix (http://bionanomatrix.com/) (43), and Complete Genomics company (http://www.completegenomics.com/) (44).

2.2.2.4. Small RNA Sequencing

The small RNA fraction is prepared for Illumina sequencing by the ligation of 5′ and 3′ RNA adapters according to Illumina's small RNA protocol, which can be found in the link http://www.illumina.com/downloads/rnaDGESmallRNA_Datasheet.pdf (45). Illumina's small RNA adaptors are ligated to the 5′ and 3′ ends of size selected <30 nt RNA. Adapter-modified DNA fragments will be enriched by PCR and further gel purified prior to sequencing. Small RNA sequencing for each sample is then performed using the Illumina Genome Analyzer (GA-2) according to the manufacturer's small RNA protocol. Typically, this protocol results in over 5–10 million small RNA sequence reads per sample per lane.

2.2.2.5. Bioinformatics Platform for Analyzing Small RNA Sequence Reads

A number of high-throughput computational pipeline have been developed for analyzing small RNA sequence reads generated by NGS technologies including Illumina sequencing (46). Our pipeline is described in (46, 47). For each sample, all unique sequence reads with a minimum read count of 10 are aligned to a reference set of miRNAs. The reference set is adaptable and currently consists of the 678 human and 472 mouse mature miRNA

sequences found in the miRNA database (miRBase version 11.0) plus 227 miRNA predictions from Berezikov et al. (48). It has been observed that the flexibility of DICER processing of the precursor miRNA produces a variety of sequence fragments, which may be active (49). To account for this, we perform a local Smith–Waterman alignment of each unique sequence read against each of the mature miRNAs in the reference, allowing for a 3-base overhang on the 5′ end and a 6-base overhang on the 3′ end. The alignments are scored such that a matching or overhanging base counts as two points and mismatches as –1. Each unique sequence read, which achieves a per-base alignment score of 2 (i.e., a perfect match) is associated with each mature miRNA for which it achieved that score. The read counts of all redundantly aligning reads are equally apportioned to all mature miRNAs to which they align.

2.2.2.6. Identification of Novel MicroRNAs

Each specimen is expected to generate multiple sequences that are not sufficiently similar to any known human miRNA. For this purpose, a number of algorithms have been developed to evaluate the likelihood that the unique sequence that does not align with a known miRNA is a putative novel miRNA. Our novel miRNA discovery pipeline is described in Creighton et al. (8, 47). First, all small RNA sequences that do not align with known miRNA precursors are mapped to the reference genome sequence of the species the small RNA is derived from (i.e., human, mouse, etc.). Each exact sequence match is fetched along with 100 bases flanking either side. These ~220-bp sequences are then tested for miRNA-like hairpin structure. The ~220-bp putative precursor sequences are evaluated with the Vienna package (www.tbi.univie.ac.at/RNA/) (50). Each of the unique sequences that map to a larger hairpin structures is tested for the Ambros criteria, which states that "authentic" miRNA sequences must map to one arm of a single-loop hairpin with a minimum free energy less than –25 kcal/mol (51, 52). Hairpins with overly large or unbalanced loops and unique sequences that map to the loop of the hairpin are rejected. After folding the read plus flanking sequence, the sequence is trimmed down to include only the plausible precursor and then folded again to ensure that the precursor was not artificially stabilized by neighboring sequence. Sequences appropriately placed in miRNA-like hairpins are considered to be "putative mature miRNAs" (pmms). Strong conservation of the mature miRNA, significant (but possibly weaker) conservation of the hairpin arm opposite the mature miRNA, and little or no conservation of the hairpin loop are considered a positive sign. Poorly conserved sequences are also considered since not all known miRNAs are conserved. If both the mature miRNA sequence and the miRNA-star sequence are found among the sequences, this candidate is

considered a definitive confirmed novel miRNA. If there is a substantial difference in abundance, the more abundant form is defined to be the mature miRNA and the less abundant form the miRNA-star sequence.

2.3. miRNA Target Prediction Programs

miRNA:mRNA target predictions for a number of different species are now available in public Web sites. We recommend the PicTar algorithm (http://pictar.bio.nyu.edu) (53), which uses predictions from Krek et al. (54); TargetScan algorithm (http://www.targetscan.org) (55), which uses predictions from Lewis et al. (56); and the miRanda algorithm (http://www.microrna.org) (57). The Sigterms software currently uses all three algorithms, and CORNA software is adapted for miRanda predictions (58).

Currently available algorithms are diverse, both in approach and in performance and all have room for improvement (7). A comparative description of some of the better known algorithms and their features are given below.

2.3.1. TargetScan (http://www.targetscan.org/)

TargetScan provides target predictions for mammalian/vertebrates offering predictions with site conservation consideration as well as without site conservation consideration (55). In predicting targets, the algorithm takes into account parameters such as stringent seed pairing, site number and factors influencing site accessibility. In the mode where site conservation is taken into account, there is an option to rank by preferential conservation instead of site context (7).

2.3.2. PicTar (http://pictar.mdc-berlin.de/)

PicTar provides target predictions for a wider variety of clades including mammalian/vertebrate, fly, and worm (53). The factors taken into consideration in this algorithm are stringent seed pairing for at least one of the sites of the miRNA, site number, and overall pairing stability (59). PicTar takes into consideration site conservation for all cases and does not offer a feature where target predictions can be done without taking conservation into account (7).

2.3.3. miRanda (http://www.microrna.org)

The miRanda algorithm is capable of making miRNA target predictions for mammal/vertebrate, fly, worm as well as additional species (56). In its criteria for target prediction, the algorithm takes into account site number, pairing to most of the miRNA, and moderately stringent seed paring (60).

2.3.4. PITA (http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html)

The PITA algorithm is capable of predicting miRNA–mRNA targets for mammalian/vertebrate, fly, and worm clades with site conservation consideration as well as without site conservation consideration (61). In its model for target predictions, PITA uses predicted site accessibility and stability as well as moderately stringent base paring and the number of sites (62).

3. Methods

Methods and software for integrating gene expression results with miRNA expression can help to maximally assess the role of miRNAs as integrators of genes into biologically meaningful networks. This is based on the fact that a given miRNA typically has predicted target sites in the 3′-UTRs of hundreds of genes and a given mRNA has multiple binding sites for several different miRNAs. In addition, genes that belong to specific pathways or networks are coordinately regulated. For all these reasons, it is essential that miRNA–mRNA association analyses are dealt with in the context of genome-wide changes in transcripts. The ultimate aim is to determine predicted miRNA–mRNA pairs that are correlated in expression in the context of a specific experiment. These could be the genes which are significantly differentially expressed when comparing two different biological states or genes that remain correlated in a treatment time course.

Current insight suggests that miRNAs exert their biologic effects by posttranscriptionally targeting gene expression; it follows that low expression of a given miRNA in a given system should conceivably cause a concomitant reversal of expression patterns for in silico predicted gene targets. Given this, we could define a miRNA–mRNA *functional* pair as consisting of a miRNA being predicted to interact with a given mRNA, where the two are also anticorrelated with each other in terms of expression. Public gene targeting prediction databases usually provide Web interface, where the user can look up predicted miRNA–mRNA functional pairs for a specific miRNA or gene of interest. In cases where the number of genes of interest (e.g., a set of genes arising from an expression profiling experiment) is in the hundreds, a gene-by-gene approach to looking up miRNA–mRNA pairs becomes impractical. Below we describe public software tools designed to make the task of integrating lists of genes and miRNAs easier.

3.1. CORNA (http://corna.sf.net)

CORNA (63) is an open-source package for the free statistical software R (http://www.r-project.org) (64) and allows scientists to analyze gene lists in the context of miRNA target predictions. In particular, when a list of genes and a list of miRNA target predictions are given, CORNA will carry out enrichment analysis to determine whether the gene list is enriched for particular miRNA targets more than that can be expected by chance. For example, the input gene list can come from a significant gene list from a microarray experiment or a biological pathway. Further methods within CORNA exist to test for significant associations

between miRNAs, pathways, and gene ontology (GO) terms and to display quantitative data associated with miRNA targets. CORNA employs three complementary statistical methods for enrichments analysis of relationships within lists of genes: the HyperGeometric test, Fisher's exact test, and the χ^2-test. Central to the flow of information through CORNA is the gene list, from which the user may test for significant miRNA target associations. The user may also start with a miRNA, find genes that are targeted by that miRNA, and then test that gene list for enrichment of KEGG pathways or GO terms. The user may also plot quantitative data associated with the targets of a particular miRNA. CORNA exclusively uses R vectors and data frames and includes functions for reading miRNA target data directly from miRBase (65) and microRNA.org (60). There are also helper functions to read gene and GO term data using biomaRt (66); microarray data directly from GEO (67); and pathway data directly from KEGG (68). A comprehensive tutorial exists at http://corna.sf.net.

3.2. Sigterms (http://sigterms.sourceforge.net)

Like CORNA, the Sigterms package allows the user to obtain miRNA–mRNA relationships for an entire set of genes (69). While CORNA runs with R, Sigterms consists of a set of Excel macros. The user enters a set of selected genes into an Excel "Annotation" workbook, which represents the entire set of genes on the gene profiling platform. The Annotation workbook can contain miRNA target predictions from one of the three commonly used algorithms (TargetScan, PicTar, and miRanda), as well GO annotation or other pathway information. Annotation workbooks for a given gene array platform representing human or mouse genes can be found at http://sigterms.sourceforge.net.

The user-provided list of genes is first entered into a Microsoft Excel document. The software will then look up the genes in the Annotation workbook to retrieve all miRNA–mRNA pairs for the given algorithm. For each miRNA, Sigterms computes an enrichment statistic that determines if the set of genes that are differentially expressed in the context of an experiment have binding sites more than expected by chance for that particular miRNA. Sigterms outputs the entire set of miRNA–mRNA pairs into an Excel worksheet; the user can then filter this worksheet for the *miRNAs* of interest (e.g., those miRNAs that are anticorrelated in expression with the genes). For computing the one-sided Fisher's exact tests for enrichment of a set of targets for a particular miRNA within the set of genes, the reference gene set determined by the complete probe set on a given array is used. To account for multiple testing of miRNAs, Monte Carlo simulation testing is performed using a 100 randomly generated gene sets. For a given gene set and a given target

prediction database, the number of miRNAs having a nominal significant p-value ($p < 0.05$) for target enrichment is computed for each of the 100 random tests. To calculate FDR, the average number of miRNA associations less than or equal to the given nominal p-value for the 100 random tests is used. The ultimate goal is to identify predicted targets within the gene set, that are enriched or overrepresented, which could help to implicate roles for specific miRNAs and miRNA-regulated genes in the system under study.

3.3. MMIA (http://129.79.233.81/~MMIA/mmia_main.html)

MMIA (which stands for "MicroRNA and mRNA integrated analysis") is a Web-based application meant to provide a "one-stop" combined analysis of the miRNA/mRNA input data for various pathway-associated gene sets (70). The user inputs mRNA expression data as a tab-delimited text file along with either a miRNA expression data table or a list of top expressed miRNAs. Given the user-defined statistical cutoff values, MMIA defines the differentially expressed genes and miRNAs from the data. Using miRNA prediction algorithms (TargetScan, PITA, and PicTar), MMIA then matches the upregulated or overexpressed genes with the downregulated or underexpressed miRNAs, and vice versa. MMIA can also generate heat maps of the data and search mRNA–miRNA pairs for pathway-related gene set enrichment. The MMIA software offers a convenient way for users to upload and analyze their data, though less flexible in how the analysis is carried out, as compared to CORNA or Sigterms.

Programs such as CORNA, Sigterms, and MMIA provide investigators without substantial bioinformatics support means by which they could make optimal use of their gene and miRNA expression data. The aim is to generate a list of miRNA–mRNA associations that are significantly correlated in the experiment of interest. The goal is to provide short lists of miRNA–mRNA pairs to be validated by direct biochemical assays, which establish that the miRNA–mRNA pair occurs in a duplex and coimmunoprecipitates in Argonaute complexes (71) and functional assays that demonstrate that the 3′-UTR of the mRNA is responsive to the cognate miRNA in luciferase or GFP reporter systems (72).

Acknowledgments

PHG and JBT are supported by a 1 R01 HL095382-01 grant. The authors would like to thank Gayani Rajapakse, Ana Hernandez, and Rajib Ghosh at University of Houston, Department of Biology and Biochemistry for their assistance in preparing this manuscript.

References

1. Freidman, R.C., Farh, K.K., Burge, C.B., and Bartel, D. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92–105.
2. Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanisms, and function. *Cell* **116**(2), 281–97.
3. Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. (2004) Human Argonaute 2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* **15**, 185–97.
4. Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO* **20**(23), 6877–88.
5. Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305**, 1434–7.
6. Olsen, P. H. and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol* **216**, 671–80.
7. Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell* **136**(2), 215–33. Review.
8. Creighton, C.J., Reid, J.G., and Gunaratne, P.H. (2009) Expression profiling of microRNAs by deep sequencing. *Brief Bioinform* **10**(5), 490–97.
9. Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–70.
10. Pariset, L., Chillemi, G., Bongiorni, S., Spica, V.R., and Valentini, A. (2009) Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. *N Biotechnol* **25**(5), 272–9.
11. NCBI microarray data repository Gene Expression Omnibus (Geo), Accessed on August 26, 2009 at http://www.ncbi.nlm.nih.gov/geo/.
12. Affymetrix GeneChip, Accessed on August 27, 2009 at http://www.affymetrix.com.
13. Agilent chip, Accessed on August 31, 2007 at http://www.agilent.com.
14. Nimblegen, Accessed on August 28, 2009 at http://www.nimblegen.com/.
15. CombiMatrix, Accessed on August 28, 2009 at http://www.combimatrix.com.
16. Illumina bead array, Accessed on August 27, 2009 at http://www.illumina.com.
17. Fan, J.B., Oliphant, A., Shen, R. et al. (2003) Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol* **68**, 69–78.
18. Shingara, J., Keiger, K., Shelton, J. et al. (2005) An optimized isolation and labeling platform for accurate microRNA expression profiling. *RNA* **11**, 1461–70.
19. Yin, J.Q., Zhao, R.C., and Morris, K.V. (2008) Profiling microRNA expression with microarrays. *Trends Biotechnol* **26**(2), 70–6.
20. Lockhart, D.J., Brown, E.L., Wong, G.G., Chee, M.S., and Gingeras, T.R. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675–80.
21. Prosnyak, M.I., Veselovskaya, S.I., Myasnikov, V.A. et al. (1994) Substitution of 2-aminoadenine and 5-methylcytosine for adenine and cytosine in hybridization probes increases the sensitivity of DNA fingerprinting. *Genomics* **21**, 490–94.
22. Rampal, J.B., ed. (2001) DNA arrays: methods and protocols. In *Methods in Molecular Biology*, Vol. 170. Humana Press, Totowa.
23. Kumar, P., Singh, S.K., Koshkin, A.A., Rajwanshi, V.K., Meldgaard, M., and Wengel, J. (1998) The first analogues of LNA (locked nucleic acids): phosphorothioate-LNA and 2′-thio-LNA. *Bioorg Med Chem Lett* **8**(16), 2219–22.
24. Arora, A., Kaur, H., Wenger, J., and Maiti, S. (2008) Effect of locked nucleic acid (LNA) modification on hybridization kinetics of DNA duplex. *Nucleic Acids Symp Ser* **52**, 417–18.
25. Agilent human, mouse, and rat miRNA microarrays product note. Agilent Technologies, Accessed on August 27, 2009 at http://www.home.agilent.com.
26. Exiqon LNA microarrays, Accessed on August 29, 2009 at http://www.exiqon.com.
27. Invitrogen, Accessed on August 28, 2009 at http://www.invitrogen.com.
28. LC Sciences, Accessed on August 30, 2009 at http://www.lcsciences.com.
29. Gao, X., Le Proust, E., Zhang, H. et al. (2001) Flexible DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res* **29**, 4744–50.

30. The institute for genomic research, Accessed on August 29, 2004 at http://www.tigr.org/.
31. Sanger, F., NIcklen, S., and Coulson, S. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–67.
32. Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560–4.
33. Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. *N Biotechnol* **25**(4), 195–203. Review.
34. The 454 genome sequencer FLX instrument (Roche Applied Science), Accessed on August 28, 2009 at http://www.454.com/.
35. Mardis, E.R. (2008) Next-generation DNA sequencing. *Annu Rev Genomics Hum Genet* **8**, 387–402.
36. The Illumina (Solexa) Genome Analyzer, Accessed on August 30, 2009 at http://www.illumina.com/.
37. The Applied Biosystems ABI SOLiD system, Accessed on August 30, 2009 at http://www3.appliedbiosystems.com/.
38. Helicos Biosciences Corporation, Accessed on August 28, 2009 at http://www.helicosbio.com/.
39. Visigen Biotechnologies Inc, Accessed on August 30, 2009 at http://visigenbio.com/.
40. Pacific Biosciences, Accessed on August 28, 2009 at http://www.pacificbiosciences.com/index.php.
41. Sequenom Inc, Accessed on August 28, 2009 at http://www.sequenom.com.
42. Oxford Nanopore Technologies, UK, Accessed on August 28, 2009 at http://www.nanoporetech.com/.
43. BioNanomatrix, Accessed on August 28, 2009 at http://bionanomatrix.com/.
44. Complete Genomics company, Accessed on August 28, 2009 at http://www.completegenomics.com/.
45. Illumina Inc, Accessed on August 28, 2009 at http://www.illumina.com/downloads/rnaDGESmallRNA_Datasheet.pdf.
46. Reid, J.G., Nagaraja, A.K., Lynn, F.C. et al. (2008) Mouse let-7 miRNA populations exhibit RNA editing that is constrained in the 5′-seed/cleavage/anchor regions and stabilize predicted mmu-let-7a:mRNA duplexes. *Genome Res* **18**(10), 1571–81.
47. Creighton, C.J., Nagaraja, A.K., Hanash, S.M., Matzuk, M.M., and Gunaratne, P.H. (2008) A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *RNA* **14**(11), 2290–6.
48. Berezikov, E., Plasterk, R.H.A., and Cuppen, E. (2002) GENOTRACE: cDNA-based local GENOme assembly from TRACE archives. *Bioinformatics* **18**(10), 1396–97.
49. Zamore, P. and Du, T. (2005) microPrimer: the biogenesis and function of microRNA. *Development* **132**, 4645–52.
50. Vienna package, Accessed on August 29, 2009 at http://www.tbi.univie.ac.at/RNA/.
51. Ambros, V., Bartel, B., Bartel, D.P. et al. (2003) A uniform system for microRNA annotation. *RNA* **9**(3), 277–9.
52. Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911–40.
53. PicTar algorithm, Accessed on August 27, 2009 at http://pictar.bio.nyu.edu.
54. Krek, A., Grun, D., Poy, M.N. et al. (2005) Combinatorial microRNA target predictions. *Nat Genet* **37**, 495–500.
55. TargetScan algorithm, Accessed on August 29, 2009 at http://www.targetscan.org.
56. Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**(1), 15–20.
57. miRanda algorithm, Accessed on August 28, 2009 at http://www.microrna.org.
58. John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. (2004) microRNA target detection. *PLoS Biol* **2**(11), 363.
59. Lall, S., Grun, D., Krek, A. et al. (2006). A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr Biol* **16**, 460–71.
60. Betel, D., Wilson, M., Gabow, A., Marks, D.S., and Sander, C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res* **36**, 149–53.
61. The PITA algorithm, Accessed on August 30, 2009 at http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html.
62. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet* **39**, 1278–84.

63. Wu, X. and Watson, M. (2009) CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics* **25**(6), 832–3.
64. The R project for statistical computing, Accessed on August, 27, 2009 at http://www.r-project.org.
65. Griffiths-Jones, S., Saini, H.K., Dongen, S.V., and Enright, A.J. (2006) miRBase: tools for micro RNA genomics. *Nucleic Acid Res* **36**, 154–8.
66. Durinck, S., Morean, Y., Kasprzyk, A. et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**(16), 3439–40.
67. Barrett, T., Suzek, T.C., Troup, D.B. et al. (2008) NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Res* **33**, 562–6.
68. Kanehisa, M., Araki, M., Goto, S. et al. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, 480–4.
69. Sigterms, Accessed on August 28, 2009 at http://sigterms.sourceforge.net.
70. miRNA and mRNA Integrated Analysis (MMIA), Accessed on August 28, 2009 at http://129.79.233.81/~MMIA/mmia_main.html.
71. Karginov, F.V., Conaco, C., Xuan, Z. et al. (2007) A biochemical approach to identifying microRNA targets. *Proc Natl Acad Sci U S A* **104**(49), 19291–6.
72. Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007) MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat Methods* **4**(9), 721–6.

LETTERS

The genome of a songbird

Wesley C. Warren[1], David F. Clayton[2], Hans Ellegren[3], Arthur P. Arnold[4], LaDeana W. Hillier[1], Axel Künstner[3], Steve Searle[5], Simon White[5], Albert J. Vilella[6], Susan Fairley[5], Andreas Heger[7], Lesheng Kong[7], Chris P. Ponting[7], Erich D. Jarvis[8], Claudio V. Mello[9], Pat Minx[1], Peter Lovell[9], Tarciso A. F. Velho[9], Margaret Ferris[2], Christopher N. Balakrishnan[2], Saurabh Sinha[2], Charles Blatti[2], Sarah E. London[2], Yun Li[2], Ya-Chi Lin[2], Julia George[2], Jonathan Sweedler[2], Bruce Southey[2], Preethi Gunaratne[10], Michael Watson[11], Kiwoong Nam[3], Niclas Backström[3], Linnea Smeds[3], Benoit Nabholz[3], Yuichiro Itoh[4], Osceola Whitney[8], Andreas R. Pfenning[8], Jason Howard[8], Martin Völker[11], Bejamin M. Skinner[12], Darren K. Griffin[12], Liang Ye[1], William M. McLaren[6], Paul Flicek[6], Victor Quesada[13], Gloria Velasco[13], Carlos Lopez-Otin[13], Xose S. Puente[13], Tsviya Olender[14], Doron Lancet[14], Arian F. A. Smit[15], Robert Hubley[15], Miriam K. Konkel[16], Jerilyn A. Walker[16], Mark A. Batzer[16], Wanjun Gu[17], David D. Pollock[17], Lin Chen[18], Ze Cheng[18], Evan E. Eichler[18], Jessica Stapley[18], Jon Slate[19], Robert Ekblom[19], Tim Birkhead[19], Terry Burke[19], David Burt[20], Constance Scharff[21], Iris Adam[21], Hugues Richard[22], Marc Sultan[22], Alexey Soldatov[22], Hans Lehrach[22], Scott V. Edwards[23], Shiaw-Pyng Yang[24], XiaoChing Li[25], Tina Graves[1], Lucinda Fulton[1], Joanne Nelson[1], Asif Chinwalla[1], Shunfeng Hou[1], Elaine R. Mardis[1] & Richard K. Wilson[1]

The zebra finch is an important model organism in several fields[1,2] with unique relevance to human neuroscience[3,4]. Like other songbirds, the zebra finch communicates through learned vocalizations, an ability otherwise documented only in humans and a few other animals and lacking in the chicken[5]—the only bird with a sequenced genome until now[6]. Here we present a structural, functional and comparative analysis of the genome sequence of the zebra finch (*Taeniopygia guttata*), which is a songbird belonging to the large avian order Passeriformes[7]. We find that the overall structures of the genomes are similar in zebra finch and chicken, but they differ in many intrachromosomal rearrangements, lineage-specific gene family expansions, the number of long-terminal-repeat-based retrotransposons, and mechanisms of sex chromosome dosage compensation. We show that song behaviour engages gene regulatory networks in the zebra finch brain, altering the expression of long non-coding RNAs, microRNAs, transcription factors and their targets. We also show evidence for rapid molecular evolution in the songbird lineage of genes that are regulated during song experience. These results indicate an active involvement of the genome in neural processes underlying vocal communication and identify potential genetic substrates for the evolution and regulation of this behaviour.

As in all songbirds, singing in the zebra finch is under the control of a discrete neural circuit that includes several dedicated centres in the forebrain termed the 'song control nuclei' (for an extensive series of reviews see ref. 8). Neurophysiological studies in these nuclei during singing have yielded some of the most illuminating examples of how vocalizations are encoded in the motor system of a vertebrate brain[9,10]. In the zebra finch, these nuclei develop more fully in the male than in the female (who does not sing), and they change markedly in size and organization during the juvenile period when the male learns to sing[11]. Analysis of the underlying cellular mechanisms of plasticity led to the unexpected discovery of neurogenesis in adult songbirds and life-long replacement of neurons[12]. Sex steroid hormones also contribute to songbird neural plasticity, in part by influencing the survival of new neurons[13]. Some of these effects are probably caused by oestrogen and/or testosterone synthesized within the brain itself rather than just in the gonads[14].

Song perception and memory also involve auditory centres that are present in both sexes, and the mere experience of hearing a song activates gene expression in these auditory centres[15]. The gene response itself changes as a song becomes familiar over the course of a day[16] or as the context of the experience changes[17]. The act of singing induces gene expression in the male song control nuclei, and these patterns of gene activation also vary with the context of the experience[18]. The function of this changing genomic activity is not yet understood, but it may support or suppress learning and help integrate information over periods of hours to days[19].

The chicken genome is the only other bird genome analysed to date[6]. The chicken and zebra finch lineages diverged about 100 million years ago near the base of the avian radiation[7]. By comparing their genomes we can now discern features that are shared (and thus

[1]The Genome Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. [2]University of Illinois, Urbana-Champaign, Illinois 61801 USA. [3]Uppsala University, Institute for Evolution and Genetics Systems, Norbyvägen 18D 752 36 Uppsala, Sweden. [4]University of California- Los Angeles, Los Angeles, California 90056, USA. [5]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [6]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [7]MRC Functional Genomics Unit, University of Oxford, Department of Physiology, Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, UK. [8]Howard Hughes Medical Institute, Department of Neurobiology, Box 3209, Duke University Medical Center, Durham, North Carolina 27710, USA. [9]Department of Behavioral Neuroscience, Oregon Health & Science University, Portland, Oregon 97239, USA. [10]Department of Biology & Biochemistry, University of Houston, Houston, Texas 77204, USA. [11]Department of Bioinformatics, Institute for Animal Health, Compton Berks RG20 7NN, UK. [12]Department of Biosciences, University of Kent, Canterbury, Kent CT2 7NJ, UK. [13]Instituto Universitario de Oncologia, Departamento de Bioquimica y Biologia Molecular, Universidad de Oviedo, 33006-Oviedo, Spain. [14]Crown Human Genome Center, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel. [15]Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904, USA. [16]Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, Louisiana 70803, USA. [17]Department of Biochemistry & Molecular Genetics, University of Colorado Health Sciences Center, Mail Stop 8101, Aurora, Colorado 80045, USA. [18]University of Washington, Genome Sciences, Seattle, Washington 98195, USA. [19]Department of Animal & Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK. [20]The Roslin Institute and Royal (Dick) School of Veterinary Studies, Edinburgh University, EH25 9OS, UK. [21]Freie Universitaet Berlin, Institut Biology, Takustr.6, 14195 Berlin, Germany. [22]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, IhnestraBe 73 14195 Berlin, Germany. [23]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [24]Monsanto Company, 800 North Lindbergh Boulevard, St Louis, Missouri 63167, USA. [25]Neuroscience Center, Louisiana State University Health Sciences Center, New Orleans, Louisiana 70112, USA.

generally characteristic of birds), and features that are most conspicuously different between the two lineages—some of which will be related to the distinctive neural and behavioural traits of songbirds.

We sequenced and assembled a male zebra finch genome using methods described previously[6,20]. A male (the homogametic sex in birds) was chosen to maximize coverage of the Z chromosome. Of the 1.2 gigabase (Gb) draft assembly, 1.0 Gb has been assigned to 33 chromosomes and three linkage groups, by using zebra finch genetic linkage[21] and bacterial artificial chromosome (BAC) fingerprint maps. The genome assembly is of sufficient quality for the analysis presented here (see Supplementary Note 1 and Supplementary Table 1). A total of 17,475 protein-coding genes were predicted from the zebra finch genome assembly using the Ensembl pipeline supplemented by Gpipe gene models (Supplementary Note 1). To extend further the characterization of genes relevant to brain and behaviour, we also sequenced complementary DNAs from the forebrain of zebra finches at 50 (juvenile, during the critical song learning period) and 850 (adult) days post-hatch, mapping these reads (Illumina GA2) to the protein-coding models (Supplementary Note 1). Of the 17,475 protein-coding gene models we find 9,872 (56%) and 10,106 (57%) genes expressed in the forebrain at these two ages (90.7% overlap), respectively. In addition to evidence for developmental regulation, these reads show further splice forms, new exons and untranslated sequences (Supplementary Figs 1 and 2).

To address issues of large-scale genome structure and evolution, we compared the chromosomes of zebra finch and chicken using both sequence alignment and fluorescent *in situ* hybridization. These analyses showed overall conservation of synteny and karyotype in the two species, although the rate of intrachromosomal rearrangement was high (Supplementary Note 2). We were also surprised to see genes of the major histocompatibility complex (MHC) dispersed across several chromosomes in the zebra finch, in contrast to the syntenic organization of both chicken and human MHCs (Supplementary Note 2).

We assessed specific gene losses and expansions in the zebra finch lineage by constructing phylogenies of genes present in the last common ancestor of birds and mammals (Supplementary Note 2 and Supplementary Fig. 3). Both the zebra finch and the chicken genome assemblies lack genes encoding vomeronasal receptors, casein milk proteins, salivary-associated proteins and enamel proteins—not surprisingly, as birds lack vomeronasal organs, mammary glands and teeth. Unexpectedly, both species lack the gene for the neuronal protein synapsin 1 (*SYN1*); comparative analyses suggest that the loss of *SYN1* and flanking genes probably occurred in an ancestor to modern birds, possibly within the dinosaur lineage (Supplementary Note 2, Supplementary Table 2 and Supplementary Fig. 4). Both zebra finch and chicken have extensive repertoires of olfactory receptor-like sequences (Supplementary Note 2 and Supplementary Fig. 5), proteases (Supplementary Table 3), and a rich repertoire of neuropeptide and pro-hormone genes.

Compared to mammals, zebra finch has duplications of genes encoding several proteins with known neural functions, including growth hormone, (Supplementary Fig. 3), caspase-3 and β-secretase (Supplementary Table 3). Two large expansions of gene families expressed in the brain seem to have occurred in the zebra finch lineage after the split from mammals. One involves a family related to the *PAK3* (p21-activated kinase) gene. Thirty-one uninterrupted *PAK3*-like sequences have been identified in the zebra finch genome, of which 29 are expressed in testis and/or brain (Supplementary Note 2). The second involves the *PHF7* gene, which encodes a zinc-finger-containing transcriptional control protein. Humans only have a single *PHF7* gene, but remarkably the gene has been duplicated independently, many times in both the zebra finch and chicken lineages to form species-specific clades of 17 and 18 genes, respectively (Supplementary Fig. 6). In the zebra finch these genes are expressed in the brain (Supplementary Note 2).

An intriguing puzzle in avian genomics has been the evident lack of a chromosome-wide dosage compensation mechanism to balance the expression of genes on the Z sex chromosome, which is present in two copies in males but only one in females[22,23]. The chicken has been suspected of exerting dosage compensation on a more local level, by the non-coding RNA MHM (male hypermethylated)[24,25], to cause a characteristic variation of gene expression along the Z chromosome. The zebra finch genome assembly, however, lacks an MHM sequence, and genes adjacent to the comparable MHM chromosomal position show no special cluster of dosage compensation (Fig. 1 and Supplementary Note 2). Thus, the putative MHM-mediated mechanism of restricted Z-chromosome dosage compensation is not common to all birds. Chromosomal sex differences in the brain could have a direct role in the sex differences so evident in zebra finch neuroanatomy and singing behaviour.

In mammals, as much as half of their genomes represent interspersed repeats derived from mobile elements, whereas the interspersed repeat content of the chicken genome is only 8.5%. We find that the zebra finch genome also has a low overall interspersed repeat content (7.7%), containing a little over 200,000 mobile elements (Supplementary Tables 4 and 5). The zebra finch, however, has about three times as many retrovirus-derived long terminal repeat (LTR) element copies as the chicken, and a low copy number of short interspersed elements (SINEs), which the chicken lacks altogether. Expressed sequence tag (EST) analysis shows that mobile elements are present in about 4% of the transcripts expressed in the zebra finch brain, and some of these transcripts are regulated by song exposure (next section, Table 1). Figure 2 shows an example of an RNA that was identified in a microarray screening for genes specifically enriched in song control nuclei[26] and now seems to represent a long non-coding RNA (ncRNA) containing a CR1-like mobile element. These results indicate that further experiments investigating a possible role of mobile-element-derived repeated sequences in vocal communication are warranted.

A large portion of the genome is directly engaged by vocal communication. A recent study[27] defined distinct sets of RNAs in the

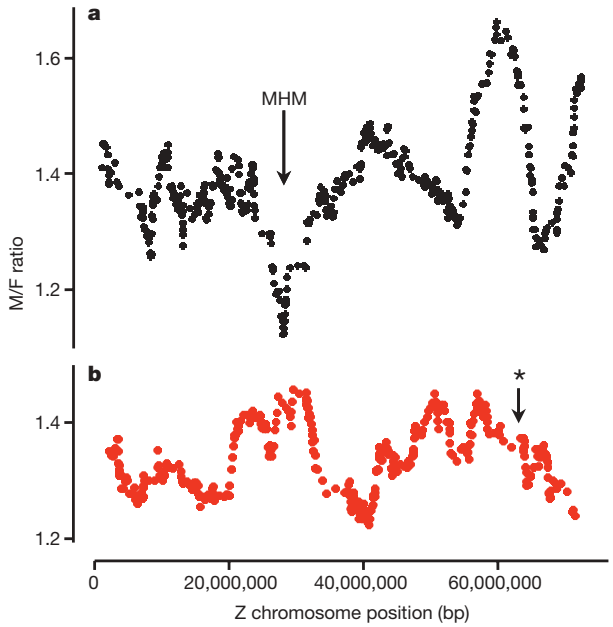


Figure 1 | Divergent patterns of dosage compensation in birds. a, **b**, The male to female (M/F) ratio of gene expression, measured by species-specific microarrays, is plotted along the Z chromosome of chicken (**a**) and zebra finch (**b**). Each point represents the average M/F ratio of a sliding window of 30 genes plotted at the median gene position and stepping one gene at a time along the chromosome. Note region of lower M/F ratios in chicken surrounding the locus of the MHM (male hypermethylated) ncRNA. In zebra finch, genes adjacent to the comparable MHM position (asterisk) show no special cluster of dosage compensation (low M/F ratios), and no MHM sequence appears in the genome assembly. bp, base pairs.

Table 1 | Structural features of the song responsive genome

	All genes analysed	Novel up	Novel down	Habituate up	Habituate down
All ESTs	17,877	145	461	1,531	1,774
Mapped loci	15,009	125	435	1,217	1,112
Ensembl genes	8,438	136	301	1,138	1,136
Mobile element content*					
Number with mobile elements	688	2	40	32	38
Percentage mobile elements	4	1	9	2	2
P-value	–	0.18	1.4×10^{-5}	0.005	0.004
Coding and non-coding content†					
mRNA transcripts (% (*P*-value))	59	86 (0.05)	32 (1×10^{-10})	65 (0.05)	71 (0.001)
EST loci mapped to introns (% (*P*-value))	6	1 (0.05)	21 (1×10^{-10})	3 (0.001)	6
Intergenic loci (% (*P*-value))	33	12 (0.001)	45 (0.05)	31	21 (0.001)
Protein-coding gene territories‡					
Mean gene length (kb)	30.4	21.7	78.8	34.8	31.2
Intergenic length (kb)	57.4	42.3	108.0	64.9	55.3
Territory size (kb)	87.8	64.1	186.8	99.7	86.4
P-value	–	3.9×10^{-3}	1.7×10^{-28}	9.3×10^{-10}	1.4×10^{-4}

A microarray made from non-redundant brain-derived ESTs[34] was used to define four subgroups of RNAs that show different responses in auditory forebrain to song exposures (novel up and down, habituated up and down)[27]. These ESTs were mapped to genome positions as described (Supplementary Note 3).
* All ESTs were analysed for mobile element content using RepeatMasker (Supplementary Note 2). *P*-value is for the comparison to all genes (Fisher's exact test).
† All ESTs that could be mapped uniquely to the genome assembly were assessed for overlap with Ensembl annotations of mRNA transcripts (protein coding and UTRs), intronic regions, or intergenic regions. *P*-value is for comparison to all mapped loci (Fisher's exact test). Results are the percentage with *P* values in parentheses where shown.
‡ The size of each unique protein-coding gene territory was determined by combining the length of the Ensembl gene model with its intergenic spacing. The *P*-value is for the comparison to all genes, using a two-tailed Wilcoxon rank sum test.

auditory forebrain that respond in different ways to song playbacks during the process of song-specific habituation, a form of learning[16]. We now map each of these song-responsive RNAs to the genome assembly (Table 1 and Supplementary Note 3). Notably, we find evidence that ~40% of transcripts in the unstimulated auditory forebrain are non-coding and derive from intronic or intergenic loci (Table 1). Among the RNAs that are rapidly suppressed in response to new vocal signals ('novel down'), two-thirds are ncRNAs.

The robust involvement of ncRNAs in the response to song led us to ask whether song exposure alters the expression of microRNAs—small ncRNAs that regulate gene expression by binding to target messenger RNAs. Indeed we find that miR-124, a conserved microRNA implicated in neurological function in other species[28], is rapidly suppressed in response to song playbacks (Fig. 3). We independently measured this effect by direct Illumina sequencing of small RNAs in the auditory forebrain, and also identified other known and new microRNAs, several of which also change in expression after song stimulation (Supplementary Note 2).

A potential site of action for microRNAs was shown by genomic mapping of transcripts that increase rapidly after new song exposure (Table 1, 'novel up'). Two of the cDNA clones that measured the most robust increases[27] align to an unusually long (3 kilobases (kb)) 3′ untranslated region (UTR) in the human gene that encodes the NR4A3 transcription factor protein (Fig. 4a). The entire UTR is similar in humans and zebra finches, with several long segments of >80% identity (Fig. 4b). Within these segments we find conserved predicted binding sites for 11 different microRNAs, including five new microRNAs found by direct sequencing of small RNAs from the zebra finch forebrain (Fig. 4b). These findings indicate that this *NR4A3* transcript element may function in both humans and songbirds to integrate many conserved microRNA regulatory pathways.

The act of singing also alters gene expression in song control nuclei[29], and we used the genome assembly to analyse the transcriptional control structure of this response. Using oligonucleotide microarrays, we identified 807 genes in which expression significantly changed as a result of singing. These were grouped by *k*-means clustering into 20 distinct expression profile clusters (Fig. 5a and Supplementary Note 3). Gene regulatory sequences (transcription-factor-binding sites) were predicted across the genome using a new motif-scanning approach (Supplementary Note 1), and we observed significant correlation between changes in expression of transcription factor genes and their predicted targets (Fig. 5b and Supplementary Table 6). Thus, the experience of singing and hearing song engages complex gene regulatory networks in the forebrain, altering the expression of microRNAs, transcription factor genes, and their targets, as well as of non-coding RNA elements that may integrate transcriptional and post-transcriptional control systems.

Learned vocal communication is crucial to the reproductive success of a songbird, and this behaviour evolved after divergence

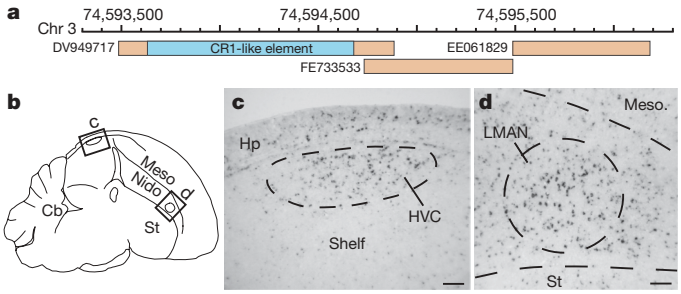


Figure 2 | Enriched expression of a CR1-like element in the zebra finch song system. **a**, Genomic alignment of an RNA containing a CR1-like retrotransposon element (in blue) and adjacent ESTs, with respective GenBank accession numbers. **b–d**, DV949717 is expressed in the brain of adult males with enrichment in song nuclei HVC (letter-based name) and LMAN (lateral magnocellular nucleus of the anterior nidopallium), as revealed by *in situ* hybridization. The diagram in **b** indicates areas shown in photomicrographs in **c** and **d**. Cb, cerebellum; Hp, hippocampus; Meso, mesopallium; Nido, nidopallium; Shelf, nidopallial shelf region; St, striatum. Scale bars, 0.1 mm.

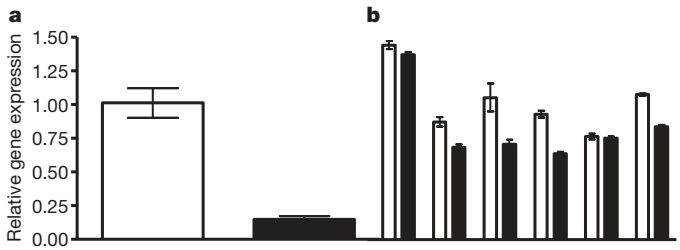


Figure 3 | miR-124 in the auditory forebrain is suppressed by exposure to new song. TaqMan assays comparing samples from the auditory lobule of adult male zebra finches in silence (open bars) or 30 min after onset of new song playback (filled bars). **a**, Comparison of two sample pools, each containing auditory forebrains of 20 birds. **b**, Comparisons of paired individual subjects, $n = 6$ pairs ($P = 0.03$, Wilcoxon paired test). Error bars denote s.e.m. of triplicate TaqMan assays. Parallel TaqMan analyses of the small RNA *RNU6B* were performed with all samples and showed no significant effect of treatment for this control RNA.

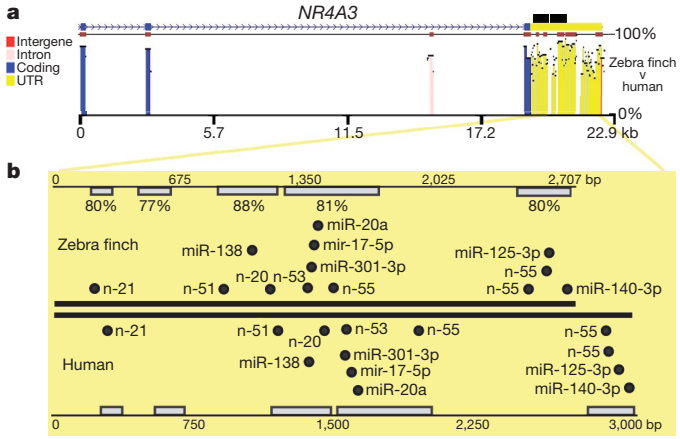


Figure 4 | Conserved *NR4A3* 3′UTR is a potential region for microRNA integration. **a**, zPicture alignment of 3′ portion of zebra finch to human gene[35] showing UTR region of high similarity beyond the coding exons. Dark red bars, regions with the highest sequence conservation; black rectangles, position of song-regulated ESTs[27] within the conserved UTR but outside the Ensembl gene model (ENSTGUG00000008853). **b**, Alignment of zebra finch and human 3′ UTR sequences showing the per cent sequence identity for each evolutionarily conserved region. Dots indicate positions of conserved new ('n-') or established ('miR-') microRNA-binding sites in both species within these regions.

of the songbird lineage[5]. Thus, it seems likely that genes involved in the neurobiology of vocal communication have been influenced by positive selection in songbirds. With this in mind, we examined the intersection of two sets of genes: (1) those that respond to song exposure in the auditory forebrain as discussed in the previous section; and (2) those that contain residues that seem to have been positively selected in the zebra finch lineage, as determined using phylogenetic analysis by maximum likelihood (PAML) (Supplementary Note 4). There are 214 genes that are common to both lists. Of these, 49 are suppressed by song exposure (Supplementary Table 7), and 6 of these 49 are explicitly annotated for ion channel activity (Table 2). This yields a highly significant statistical enrichment for the term 'ion channel activity' ($P = 0.0016$, false discovery rate (FDR) adjusted Fisher's exact test) and other related terms in this subset of genes (Supplementary Tables 8 and 9). Independent evidence has also demonstrated differential anatomical expression of ion channel genes in song control nuclei[26,30]. Ion channel genes have important roles in many aspects of behaviour, neurological function and disease[31]. This class of genes is highly likely to be linked to song behaviour and should be a major target for future functional studies.

Passerines represent one of the most successful and complex radiations of terrestrial animals[7]. Here we present the first, to our knowledge, analysis of the genome of a passerine bird. The zebra finch was chosen because of its well-developed status as a model organism for a number of fields in biology, including neurobiology, ethology, ecology, biogeography and evolution. In the zebra finch as in the chicken, we see a smaller, tighter genome compared to mammals, with a marked reduction of interspersed repeats. The zebra finch presents a picture of greater genomic plasticity than might have been expected from the chicken and other precedents, with a high degree of intrachromosomal rearrangements between the two avian species, gene copy number variations and transcribed mobile elements. Yet we also see an overall similarity to mammals in protein-coding gene content and core transcriptional control systems.

Our analysis suggests several channels through which evolution may have acted to produce the unique neurobiological properties of songbirds compared to the chicken and other animals. These include the management of sex chromosome gene expression, accelerated evolution of neuronal ion transport genes, gene duplications to produce new variants of *PHF7*, *PAK3* and other neurobiologically

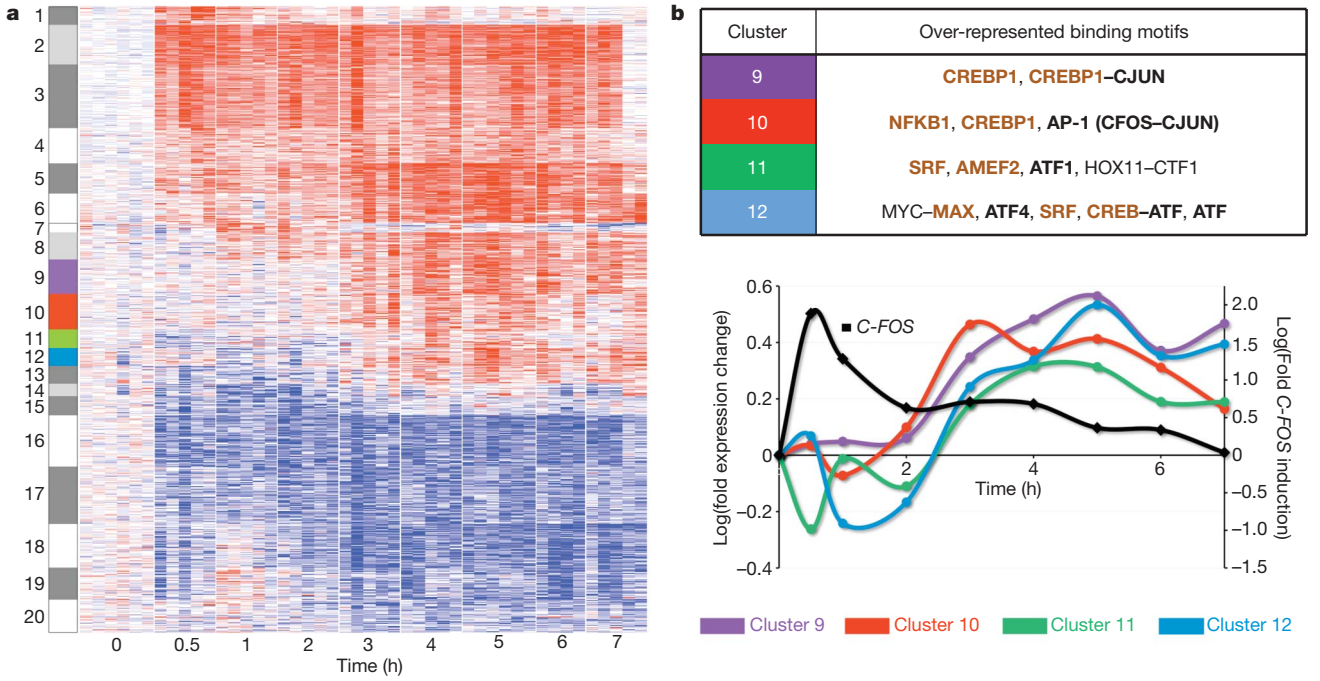


Figure 5 | Transcriptional control network in area X engaged by singing. **a**, Clustered (1–20) temporal expression profiles of 807 genes (rows) that change with time and amount of singing; red, increases; blue, decreases; white, no change relative to average 0-h control. Grey/coloured bars on left, clusters with enrichment of specific promoter motifs ($P < 0.01$). **b**, Enriched transcription-factor-binding motifs (abbreviations) found in the promoters of late response genes, clusters 9–12 (coloured as in **a**); bold, binding sites for known activity-dependent transcription factors (for example, CREBP1) or transcription factor complexes (for example, CREBP1–CJUN); black, sites for post-translationally activated transcription factors; brown, sites for transcriptionally activated transcription factors including by singing (for example, in cluster 1). Graph shows time course of average expression of all genes in the late response clusters, normalized to average 0 h for that cluster. Also plotted is the average expression of the *C-FOS* transcription factor mRNA, which binds to the AP-1 site over-represented in the promoters of cluster 10 genes.

Table 2 | Song-suppressed ion channel genes under positive selection

Gene	Description	Branch $\Delta\omega$	Sites PS/total
CACNA1B	Voltage-dependent N-type calcium channel subunit α-1B	0.016	9/2,484
CACNA1G	Voltage-dependent T-type calcium channel subunit α-1G	0.044*	2/2,468
GRIA2	Glutamate receptor 2 precursor (GluR-2, AMPA 2)	0.231*	17/948
GRIA3	Glutamate receptor 3 precursor (GluR-3, AMPA 3)	−0.010	4/894
KCNC2	Potassium voltage-gated channel subfamily C member 2 (Kv3.2)	0.315*	32/654
TRPV1	Transient receptor potential cation channel subfamily V member 1	−0.067	3/876

These six genes are suppressed by song exposure (FDR = 0.05)[27] and they show evidence of positive selection in the zebra finch relative to chicken ($P < 10^{-3}$, Supplementary Note 3). Branch $\Delta\omega$ denotes the difference in the non-synonymous to synonymous substitution ratio (dN/dS) between zebra finch and other birds (chicken and the ancestral branch leading to chicken and zebra finch). Positive values indicate that the gene is rapidly evolving, whereas negative values indicate genes evolving more slowly. Sites PS/total denotes the number of individual sites with empirical Bayes posterior probability greater than 0.95 of $\omega > 1$ (positive selection) in the finch versus the total number of residues in the protein, from branch-site model analysis implemented in PAML. Note that genes can show overall slower evolution in the branch model yet show evidence of significant positive selection at specific sites.
* Gene-wide differences that were significant ($P < 0.05$) by a likelihood ratio test.

important genes, and a new arrangement of MHC genes. Most notably, our analyses suggest a large recruitment of the genome during vocal communication, including the extensive involvement of ncRNAs. It has been proposed that ncRNAs have a contributing role in enabling or driving the evolution of greater complexity in humans and other complex eukaryotes[32]. Seeing that learned vocal communication itself is a phenomenon that has emerged only in some of the most complex organisms, perhaps ncRNAs are a nexus of this phenomenon.

Much work will be needed to establish the actual functional significance of many of these observations and to determine when they arose in avian evolution. This work can now be expedited with the recent development of a method for transgenesis in the zebra finch[33]. An important general lesson, however, is that dynamic and serendipitous aspects of the genome may have unexpected roles in the elaborate vocal communicative capabilities of songbirds.

METHODS SUMMARY

Sequence assembly. Sequenced reads were assembled and attempts were made to assign the largest contiguous blocks of sequence to chromosomes using a genetic linkage map[21], fingerprint map and synteny with the chicken genome assembly Gallus_gallus-2.1, a revised version of the original draft[6] (Supplementary Note 1).
Genes. Gene orthology assignment was performed using the EnsemblCompara GeneTrees pipeline and the OPTIC pipeline (Supplementary Note 1). Orthology rate estimation was performed with PAML (pairwise model = 0, Nssites = 0). In all cases, codon frequencies were estimated from the nucleotide composition at each codon position (F3X4 model).
Gene expression and evolution. Methods for Illumina read counting, *in situ* hybridization, TaqMan RT–PCR, microarrays, regulatory motif and evolutionary rate analyses are given in Supplementary Notes 1–4.

Received 30 September 2009; accepted 6 January 2010.

1. Zann, R. A. *The Zebra Finch: A Synthesis of Field and Laboratory Studies* (Oxford Univ, Press, 1996).
2. Clayton, D. F., Balakrishnan, C. N. & London, S. E. Integrating genomes, brain and behavior in the study of songbirds. *Curr. Biol.* **19,** R865–R873 (2009).
3. Nottebohm, F. in *Hope For a New Neurology* (ed. Nottebohm, F.) (New York Academy of Science, 1985).
4. Doupe, A. J. & Kuhl, P. K. Birdsong and human speech: common themes and mechanisms. *Annu. Rev. Neurosci.* **22,** 567–631 (1999).
5. Jarvis, E. D. Learned birdsong and the neurobiology of human language. *Ann. NY Acad. Sci.* **1016,** 749–777 (2004).
6. Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
7. Hackett, S. J. *et al.* A phylogenomic study of birds reveals their evolutionary history. *Science* **320,** 1763–1768 (2008).
8. Zeigler, H. P. & Marler, P. *Behavioral Neurobiology of Bird Song* Vol. 1016 (New York Academy of Sciences, 2004).
9. Hahnloser, R. H., Kozhevnikov, A. A. & Fee, M. S. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* **419,** 65–70 (2002).
10. Mooney, R. Neural mechanisms for learned birdsong. *Learn. Mem.* **16,** 655–669 (2009).
11. Konishi, M. & Akutagawa, E. Neuronal growth, atrophy and death in a sexually dimorphic song nucleus in the zebra finch brain. *Nature* **315,** 145–147 (1985).
12. Goldman, S. A. & Nottebohm, F. Neuronal production, migration, and differentiation in a vocal control nucleus of the adult female canary brain. *Proc. Natl Acad. Sci. USA* **80,** 2390–2394 (1983).
13. Nottebohm, F. The road we travelled: discovery, choreography, and significance of brain replaceable neurons. *Ann. NY Acad. Sci.* **1016,** 628–658 (2004).
14. London, S. E., Remage-Healey, L. & Schlinger, B. A. Neurosteroid production in the songbird brain: A re-evaluation of core principles. *Front. Neuroendocrinol.* **30,** 302–314 (2009).
15. Mello, C. V., Vicario, D. S. & Clayton, D. F. Song presentation induces gene expression in the songbird forebrain. *Proc. Natl Acad. Sci. USA* **89,** 6818–6822 (1992).
16. Dong, S. & Clayton, D. F. Habituation in songbirds. *Neurobiol. Learn. Mem.* **92,** 183–188 (2009).
17. Woolley, S. C. & Doupe, A. J. Social context-induced song variation affects female behavior and gene expression. *PLoS Biol.* **6,** e62 (2008).
18. Jarvis, E. D., Scharff, C., Grossman, M. R., Ramos, J. A. & Nottebohm, F. For whom the bird sings: context-dependent gene expression. *Neuron* **21,** 775–788 (1998).
19. Clayton, D. F. The genomic action potential. *Neurobiol. Learn. Mem.* **74,** 185–216 (2000).
20. Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453,** 175–183 (2008).
21. Stapley, J., Birkhead, T. R., Burke, T. & Slate, J. A linkage map of the zebra finch *Taeniopygia guttata* provides new insights into avian genome evolution. *Genetics* **179,** 651–667 (2008).
22. Itoh, Y. *et al.* Dosage compensation is less effective in birds than in mammals. *J. Biol.* **6,** 2 (2007).
23. Ellegren, H. *et al.* Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. *BMC Biol.* **5,** 40 (2007).
24. Teranishi, M. *et al.* Transcripts of the MHM region on the chicken Z chromosome accumulate as non-coding RNA in the nucleus of female cells adjacent to the *DMRT1* locus. *Chromosome Res.* **9,** 147–165 (2001).
25. Arnold, A. P., Itoh, Y. & Melamed, E. A bird's-eye view of sex chromosome dosage compensation. *Annu. Rev. Genomics Hum. Genet.* **9,** 109–127 (2008).
26. Lovell, P. V., Clayton, D. F., Replogle, K. L. & Mello, C. V. Birdsong "transcriptomics": neurochemical specializations of the oscine song system. *PLoS One* **3,** e3440 (2008).
27. Dong, S. *et al.* Discrete molecular states in the brain accompany changing responses to a vocal signal. *Proc. Natl Acad. Sci. USA* **106,** 11364–11369 (2009).
28. Makeyev, E. V. & Maniatis, T. Multilevel regulation of gene expression by microRNAs. *Science* **319,** 1789–1790 (2008).
29. Wada, K. *et al.* A molecular neuroethological approach for identifying and characterizing a cascade of behaviorally regulated genes. *Proc. Natl Acad. Sci. USA* **103,** 15212–15217 (2006).
30. Wada, K., Sakaguchi, H., Jarvis, E. D. & Hagiwara, M. Differential expression of glutamate receptors in avian neural pathways for learned vocalization. *J. Comp. Neurol.* **476,** 44–64 (2004).
31. Cooper, E. C. & Jan, L. Y. Ion channel genes and human neurological disease: recent progress, prospects, and challenges. *Proc. Natl Acad. Sci. USA* **96,** 4759–4766 (1999).
32. Mattick, J. S. RNA regulation: a new genetics? *Nature Rev. Genet.* **5,** 316–323 (2004).
33. Agate, R. J., Scott, B. B., Haripal, B., Lois, C. & Nottebohm, F. Transgenic songbirds offer an opportunity to develop a genetic model for vocal learning. *Proc. Natl Acad. Sci. USA* **106,** 17963–17967 (2009).
34. Replogle, K. *et al.* The Songbird Neurogenomics (SoNG) Initiative: community-based tools and strategies for study of brain gene function and evolution. *BMC Genomics* **9,** 131 (2008).
35. Ovcharenko, I., Loots, G. G., Hardison, R. C., Miller, W. & Stubbs, L. zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res.* **14,** 472–477 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The sequencing of zebra finch was funded by the National Human Genome Research Institute (NHGRI). Further research support included grants to D.F.C. (NIH RO1 NS045264 and RO1 NS051820), H.E. (Swedish Research Council and Knut and Alice Wallenberg Foundation), E.D.J. (HHMI, NIH Directors Pioneer Award and R01 DC007218), M.A.B. (NIH RO1 GM59290) and J.S. (Biotechnology and Biological Sciences Research Council grant number BBE0175091). Resources for exploring the sequence and annotation data are

available on browser displays available at UCSC (http://genome.ucsc.edu), Ensembl (http://www.ensembl.org), the NCBI (http://www.ncbi.nlm.nih.gov) and http://aviangenomes.org. We thank K. Lindblad-Toh for permission to use the green anole lizard genome assembly, the Production Sequencing Group of The Genome Center at Washington University School of Medicine for generating all the sequence reads used for genome assembly, and the Clemson University Genome Institute for the construction of the BAC library. We would like to recognize all the important published work that we were unable to cite owing to space limitations.

Author Contributions W.C.W., D.F.C., H.E. and A.P.A. comprise the organizing committee of the zebra finch genome sequencing project. Project planning, management and data analysis: W.C.W., D.F.C., H.E. and A.P.A. Assembly annotation and analysis: L.W.H., P.M., S.-P.Y., L.Y., J.N., A.C., S.H., J.Sl., J.St., D.B. and S.-P.Y. Protein coding and non-coding gene prediction: S.S., C.B., P.F., S.W., A.H., C.P.P. and L.K. SNP analysis: P.F. and W.M.M. Orthology prediction and analysis: A.J.V., A.H., C.P.P., S.F. and L.K. Repeat element analysis: M.A.B., A.F.A.S., R.H., M.K.K., J.A.W., W.G. and D.D.P. Segmental duplication and gene duplication analysis: L.C., Z.C., E.E.E., L.K., C.P.P., M.F., C.N.B., R.E., J.G. and S.E.L. Protease annotation and analysis: X.S.P., V.Q., G.V. and C.L.-O. Neuropeptide hormone annotation: J.Sw. and B.S. Small non-coding RNA analysis: Y-C.L., Y.L., P.G., M.W. and X.L. Comparative mapping: D.K.G., M.V. and B.M.S. Singing induced gene network analysis: E.D.J., A.R.P., O.W. and J.H. Z-chromosome analysis: Y.I. and A.P.A. Gene expression and *in situ* analysis and synapsin synteny/loss analysis: C.V.M., P.L. and T.A.F.V. Adaptive evolution analysis: A.K., K.N., N.B., L.S., B.N. and C.N.B. Gene expression in the brain analysis: C.S., I.A., A.S., H.L., H.R. and M.S. MHC analysis: S.E., C.N.B. and R.E. Olfactory receptor analysis: T.O., D.L. and L.K. Sequencing management: R.K.W., E.R.M. and L.F. Physical map construction: T.G. Zebra finch tissue resources: T.Bu. and T.Bi. Zebra finch cDNA resources: D.F.C., E.D.J. and X.L.

Author Information The *Taeniopygia guttata* whole-genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the project accession ABQF00000000. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.C.W. (wwarren@watson.wustl.edu), D.F.C (dclayton@illinois.edu), H.E. (hans.ellegren@ebc.uu.se) or A.P.A. (arnold@ucla.edu).

BMC Genomics

RESEARCH ARTICLE Open Access

Song exposure regulates known and novel microRNAs in the zebra finch auditory forebrain

Preethi H Gunaratne[1,2,3†], Ya-Chi Lin[4†], Ashley L Benham[1], Jenny Drnevich[5], Cristian Coarfa[11], Jayantha B Tennakoon[1], Chad J Creighton[6], Jong H Kim[1], Aleksandar Milosavljevic[11], Michael Watson[7], Sam Griffiths-Jones[8] and David F Clayton[4,9,10*]

Abstract

Background: In an important model for neuroscience, songbirds learn to discriminate songs they hear during tape-recorded playbacks, as demonstrated by song-specific habituation of both behavioral and neurogenomic responses in the auditory forebrain. We hypothesized that microRNAs (miRNAs or miRs) may participate in the changing pattern of gene expression induced by song exposure. To test this, we used massively parallel Illumina sequencing to analyse small RNAs from auditory forebrain of adult zebra finches exposed to tape-recorded birdsong or silence.

Results: In the auditory forebrain, we identified 121 known miRNAs conserved in other vertebrates. We also identified 34 novel miRNAs that do not align to human or chicken genomes. Five conserved miRNAs showed significant and consistent changes in copy number after song exposure across three biological replications of the song-silence comparison, with two increasing (tgu-miR-25, tgu-miR-192) and three decreasing (tgu-miR-92, tgu-miR-124, tgu-miR-129-5p). We also detected a locus on the Z sex chromosome that produces three different novel miRNAs, with supporting evidence from Northern blot and TaqMan qPCR assays for differential expression in males and females and in response to song playbacks. One of these, tgu-miR-2954-3p, is predicted (by TargetScan) to regulate eight song-responsive mRNAs that all have functions in cellular proliferation and neuronal differentiation.

Conclusions: The experience of hearing another bird singing alters the profile of miRNAs in the auditory forebrain of zebra finches. The response involves both known conserved miRNAs and novel miRNAs described so far only in the zebra finch, including a novel sex-linked, song-responsive miRNA. These results indicate that miRNAs are likely to contribute to the unique behavioural biology of learned song communication in songbirds.

Background

Songbirds are important models for exploring the neural and genomic mechanisms underlying vocal communication, social experience and learning (reviewed in [1]). Songbirds communicate using both innate calls and learned vocalizations (songs), and unique specializations of the brain evolved to support this behavior (reviewed in [2]). In the zebra finch, only the male produces songs, although both sexes process and discriminate specific songs [3-6]. The genome is actively engaged by song communication, as first shown in an early demonstration of how gene responses in the brain discriminate among different song stimuli [7]. The genomic response is not a simple correlate of neural activity and it can vary significantly according to the salience and behavioral context of the experience [8-13]. Recent studies using microarray technology have now shown that song exposure affects the expression of thousands of genes in the auditory forebrain [14,15]. Repeated exposure to one song leads to an altered gene expression profile, correlated with habituation of both the behavioral and immediate genomic responses to that specific song. These observations suggest the involvement of large and dynamic transcriptional network in the recognition and memory of complex vocal signals [14].

MicroRNAs (miRNAs or miRs) are emerging as potential control points in transcriptional networks, and

* Correspondence: dclayton@uiuc.edu
† Contributed equally
[4]Department of Cell and Developmental Biology, University of Illinois, Urbana-Champaign, IL 61801, USA
Full list of author information is available at the end of the article

may be particularly important for the evolution of brain and behavior. Many miRNAs are expressed in the brain [16], often in different patterns in different species [17-19]. Brain miRNAs undergo dramatic changes in expression during development [20-22] and aging [23] and have been functionally implicated in neurological disease [24]. They may also function in the normal physiological operation of the nervous system as suggested by evidence for involvement of miR-132 and miR-219 in circadian clock regulation [25] and miR-134 in control of dendritic translation [26,27].

Here we apply massively parallel Illumina sequencing to probe the involvement of miRNAs in the processing of song experience in the zebra finch auditory forebrain. We begin by identifying 155 different miRNA sequences and the genomic loci of their precursor sequences in the zebra finch genome, including 34 miRNA genes that have not been detected in the genomes of other species. We then ask whether the miRNA content changes after song exposure and find robust evidence of miRNA responses to song playbacks. We also assess correlations between expression changes of a novel miRNA and its predicted target mRNAs during song habituation. The results indicate an active role for miRNAs in the neural processing of a natural perceptual experience - hearing the sound of another bird singing.

Results

The miRNAs of the zebra finch auditory forebrain

We carried out Illumina small RNA sequencing (RNA-seq) on the small RNA (~18-30 nucleotides) fraction of total RNA isolated from adult zebra finch auditory forebrain. Ultimately, we performed 6 Illumina runs on 6 different RNA samples, to assess the effects of song exposure (next section). First we describe the overall small RNA profile obtained by combining the results of all the runs, representing 36 adult zebra finches (equal numbers of males and females). A total of 20 million reads were obtained (Table 1) and aligned to reference miRNA sequences from other species (miRBase version 13.0). Overall we identified 107 non-redundant miRNAs representing 52% of sequences that have been previously identified in chicken, rodent and human. The remaining sequences mapping to the piRNA database were denoted as piRNA reads (~30%) (Additional File 1, Table S1).

Reads that did not align to known RNAs were assessed for miRNA potential through a novel miRNA discovery pipeline described by Creighton et al.[28] which tests for properties that are characteristic of known miRNAs. These properties include the following: 1) The mature sequence must map to the stem region of the hairpin sequence of the putative precursor extracted from the zebra finch genome. 2) The mature miRNA sequence must map to the precursor such that it can be processed following the Drosha processing rules [29]. All novel miRNA candidates that map to the loop region and/or lack appropriate Drosha processing sites are failed. 3) Known miRNAs have stable 5'-ends that vary at the most by +/- 1 nucleotide. 4) By contrast the 3'-ends of miRNAs are highly heterogeneous in length due to imprecise Dicer processing [29,30] and exhibit non-templated nucleotide sequence changes due to RNA editing [29-31]. 5) Consequently, the putative precursor must give a strong signal of sequence alignments in a tight area of 18-25 nucleotides. Small RNA sequences that are distributed fairly evenly along the entire length of the precursor are rejected since they likely represent degraded products of a large RNA. The candidates that also demonstrate the presence of the miRNA star sequence (miR*) mapping on the opposite side of the mature miRNA and occurring at a lower abundance in the deep sequencing data are considered to be confirmed novel miRNAs in zebra finch. Using this pipeline (Figure 1) we discovered 48 putative novel miRNAs that map on the zebra finch genome to a stem loop structure that folds with a minimum free energy of < -20 kcal/mol [32]. The complete analysis and mapping information for all the novel miRNA candidates is given in Additional File 1, Tables S2 and S3.

All novel miRNA candidates were mapped to genomic loci in the zebra finch genome assembly [33], and also to human and chicken genomes using the BLAT function of the UCSC Genome Browser (Additional File 1, Table S3). In the zebra finch genome, the loci include both annotated exons and introns as well as unannotated intergenic regions. Thirty-four (34) novel microRNAs uncovered from zebra finch are not presently detected in the human or chicken genome assemblies. Eleven (11) map to genome positions in chicken, and six to positions in the human (with three of these found in

Table 1 Summary statistics for the read alignments

		Male silence	Male song	Female silence	Female song	Mix silence	Mix song
Total Reads		2,704,778	2,056,391	3,173,108	3,546,038	3,962,050	4,738,528
Total Usable Reads		1,179,330	1,155,168	2,244,376	2,498,648	2,249,188	2,950,398
Reads aligning with	Total	401,934	209,944	1,638,528	1,755,748	1,348,109	2,113,006
known miRNA	Fraction	34%	18%	73%	70%	60%	72%

Six different pools of auditory forebrain were analyzed independently by Illumina small RNA sequencing, as described in the text.

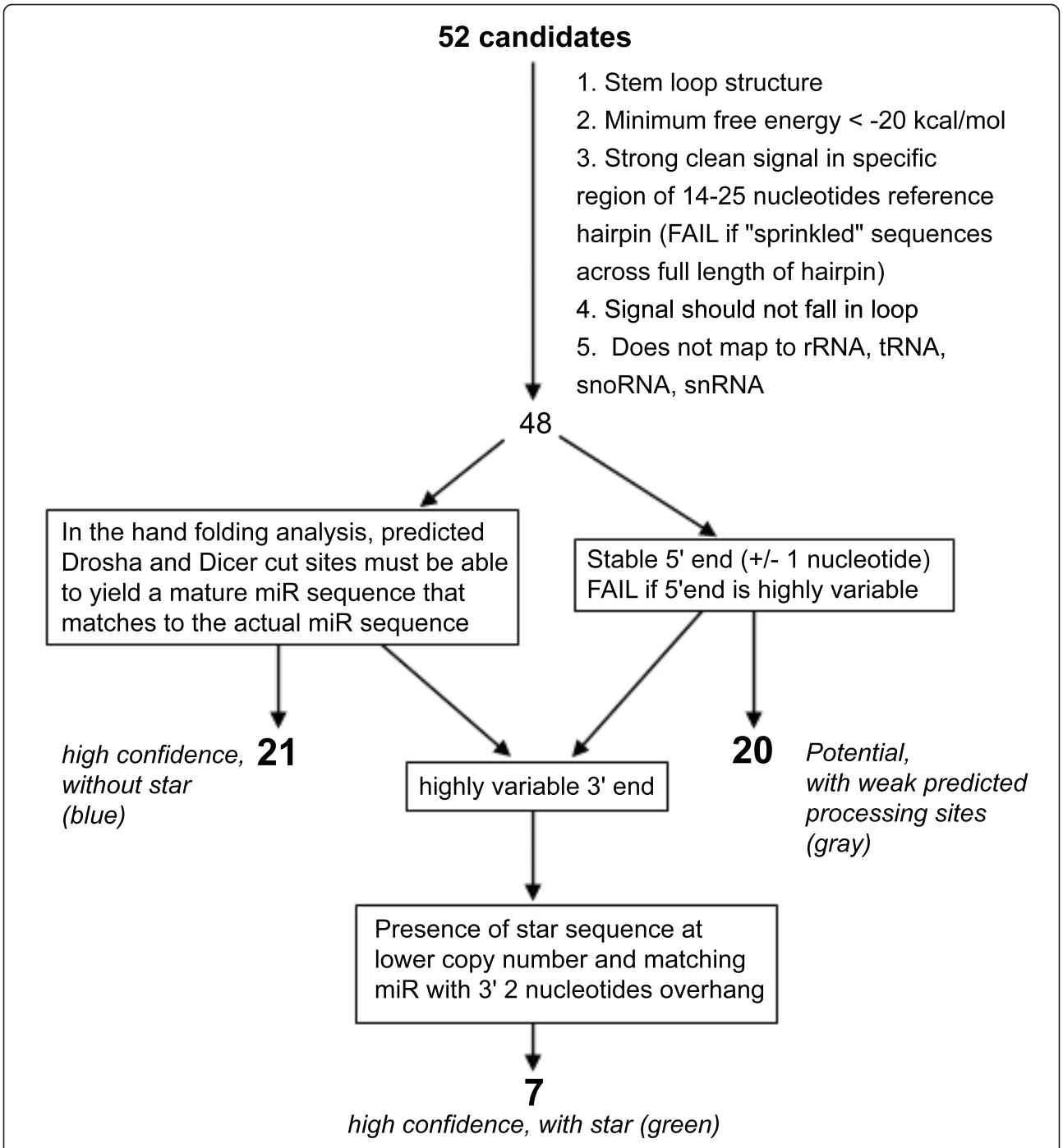


Figure 1 Pipeline with yields for analysis of putative novel miRNAs. 52 small RNA sequences did not align to miRBase reference sequences and were assessed for miRNA potential. 48 sequences passed the minimum criteria and were categorized into three groups according to strength of evidence (sequences are color-coded in Additional File 1, Table S3, as indicated). Seven (7) are confirmed novel miRNAs since they had all the characteristics of known miRNAs and in addition also had a less abundant miR* sequence that maps on the opposite side of the stem from the putative novel miRNA. These are labelled green in Additional File 1, Table S3. Twenty-one (21) putative novel miRNAs are highly confident (labelled blue) since they also shared characteristics of known miRNAs but no sequence was found aligning to the miR* region. Given that the miR and miR* sequences for most known miRNAs have vastly different copy numbers such that the miR* sequence is sometimes not found, the highly confident candidates are also highly likely to be genuine novel miRNAs, Twenty (20) candidates (labelled grey) had a subset of the characteristics of known miRNAs but not all and therefore were deemed potential candidates that require more evidence.

human but not chicken assemblies). Tgu-mir-2976 maps to three loci in the finch and 14 in the chicken, indicating a probable expansion of this miRNA in the chicken lineage. This putative novel miRNA is not currently detected in the human assembly HG18. Tgu-mir-2985 is intriguing as it is located within two stem loops within the introns of two functionally related genes: the glutamate receptor subunits GRIA2 and GRIA4 in all three genomes.

miRNA responses to song exposure

When zebra finches are exposed to playback of a song they have not heard recently, changes occur in the expression of many different mRNAs as detected 30 min after stimulus onset [14]. To determine whether specific miRNAs also change in expression, we counted the Illumina reads in samples of RNA pooled from the auditory forebrain of birds either 30 min after onset of song playback (Song group) or from matched controls (Silence group). In our first such experiment, the birds in both groups were all males (n = 6 each). The read count for each miRNA in each sample was normalized to the total number of usable reads mapped in that sample. We then calculated the ratio of the normalized count in the Song-stimulated condition compared to the Silence condition and performed a Fisher's exact test (with correction for multiple testing) to evaluate whether the ratio differed significantly from the range of expected values at a 95% confidence interval. In the initial experiment with males, 49 of the known conserved miRNAs showed a significant difference, with 28 decreasing and 21 increasing in the group exposed to song (Additional File 1, Table S4).

To address the biological reproducibility of the miRNA responses to song more broadly, we then repeated the small RNA-seq comparison two additional times using new groups of birds. In the second experiment, we used only females, and in the third we used an equal mix of males and females. In total, therefore, we performed three independent "song-silence" pairwise comparisons by small RNA-seq, with an overall sex balance but different sex ratios in each individual comparison. These second and third experiments were done six months after the first and Illumina technology had improved by this time so that we obtained twice as many read counts (Table 1) - but again we normalized to the total mapped read number in each individual sample for our statistic analyses. As in the first experiment, we again observed differential read counts for roughly a third of the miRNAs, but the identities of the miRNAs affected were somewhat different in each comparison. This is summarized graphically as a Venn diagram (Additional File 2, Figure S1), and comprehensive read count data are presented in Additional File 1, Table S4. Across all three experiments, five conserved miRNAs showed changes that were both significant and in same direction in all comparisons (Table 2). For a number of other miRNAs, including let-7f, an apparent effect of song exposure was measured in all three experiments but the direction of change was not consistent (Additional File 1, Table S4).

We performed TaqMan assays on RNA from additional birds, probing for eleven of the "significantly affected" miRNAs, and obtained fluorescent signals in PCR for ten. In nine out of ten cases, we observed the same direction of song response by TaqMan as in the small RNA-seq experiment, although the P-value by TaqMan was below 0.05 in only five cases (tgu-miR-124, tgu-miR-29a, tgu-miR-92, tgu-129-5p, and tgu-miR-2954-3p, Additional File 1, Table S4). The lack of statistical significance in the TaqMan assay for the others could reflect differences in the sensitivity and resolution of Illumina vs. TaqMan assays, or the operation of other uncontrolled factors in our experiments that lead to variability in the expression of some miRNAs.

The transcriptional response in the auditory forebrain of *zenk* and other mRNAs is specific to song relative to

Table 2 Conserved miRNAs with consistent responses to song exposure

	Male				Female				Mix			
	Silence	Song	Fold Change	FDR-P	Silence	Song	Fold Change	FDR-P	Silence	Song	Fold Change	FDR-P
Increasing												
tgu-miR-25	227	423	3.57	1.6E-27	55	212	3.60	1.4E-10	35	160	2.92	2.1E-05
tgu-miR-192	26	69	5.08	1.2E-06	36	90	2.33	5.5E-03	11	97	5.63	4.3E-06
Decreasing												
tgu-miR-92	359	100	0.53	1.1E-04	5479	5398	0.92	5.5E-03	7461	6887	0.59	6.3E-108
tgu-miR-124	24624	7056	0.55	2.1E-251	56802	46434	0.76	1.1E-206	50955	77220	0.97	1.6E-04
tgu-miR-129-5p	2020	602	0.57	4.0E-19	9778	7272	0.69	2.8E-62	12128	9284	0.49	2.6E-293

Shown are the Illumina read data for the five miRNAs that show a consistent response to song (same direction of change, significant in all three comparisons). "Song" and "Silence" list raw counts from the Illumina read analysis (Additional File 1, Table S4). "Fold Change" is the ratio of Song versus Silence read counts, after the raw counts were normalized within each run to the sum of mapped reads for that sample. Thus a value of > 1 indicates a relative increase in the group exposed to song, and < 1 indicates a decrease. "FDR-P" indicates the result of the Fisher's exact test (FDR adjusted) for this comparison. See Additional File 1, Table S4 for full list of values for all miRNAs, and associated TaqMan values for a subset of these miRNAs (measured in a different set of males and females).

non-song auditory stimuli [6,7,34,35]. To test for song-specificity of the miRNA response, we conducted a further TaqMan experiment assessing the levels of six miRNAs (tgu-miR-124, tgu-miR-92, tgu-miR-129-5p, and three miRNAs derived from the tgu-miR-2954 locus, next section), in birds who had heard either a normal song or a carefully matched non-song acoustic stimulus, "song enveloped noise" (SEN). SEN has the same amplitude envelope as the song from which it is derived but spectral content has been randomized so it does not sound like a song [34]. By TaqMan PCR, we confirmed that normal song induced a larger increase in *zenk* mRNA in these birds than did SEN (Additional File 2, Figure S3 panel D). In these same animals, normal song, but not SEN, triggered a significant decrease in the levels of tgu-miR-124, tgu-mir-129-5p, tgu-miR-92 and tgu-miR-2954-3p (Additional File 2, Figure S3 panels A-C, H). Thus we conclude that there is indeed a unique miRNA response in the auditory forebrain that is selective for song over non-song acoustic stimuli.

A complex sex-linked miRNA locus in zebra finch and other birds

The novel miRNA, tgu-mir-2954, that was detected most frequently in our Illumina assays maps to the sense strand of an intron in the XPA gene, on the Z chromosome (Figure 2A). The precursor hairpin contains reads from both arms, thus meeting our bioinformatic criteria for a confirmed miRNA (Figure 2B). By contrast to most known miRNAs, the numbers of reads from both 5' and 3' arms were found at similar copy numbers, suggesting that both arms may make functional mature miRNAs. BLAST analysis of the mir-2954 hairpin precursor sequence against the NCBI nr database identified a putative mature miRNA in chicken (gi|145279910|emb|AM691163.1|), and BLAT analysis of a collection of transcripts from crocodile and 11 other bird species [36] detected mir-2954 transcripts in 2 non-passerine species (two hummingbirds) and 3 passerine species (the American crow, the pied flycatcher, and the golden collared manakin) (Additional File 2, Figure S2). There was no BLAT hit in the crocodile, the remaining 3 non-passerine birds (Emu, budgerigar, and ringneck dove), and 3 passerine species (collared flycatcher, blue tit and Eastern phoebe). The lack of a hit does not necessarily mean absence of the gene as these datasets represent incomplete transcriptomes derived by 454 sequencing [36]. These results clarify that the sequence is not unique to the zebra finch or passerines, but may nevertheless have a restricted distribution within birds.

To validate the existence of these two miRNAs in zebra finch, we performed TaqMan analyses for both, using their reverse complements as controls. Interestingly, we got significant expression values not only for the predicted miRNAs but also for one of the reverse-complement miRNAs (tgu-miR-2954R-5p) although no significant song regulation for miR-2954R-5p was found (Additional File 2, Figure S3 panels I-J). With respect to the XPA gene within which this locus is embedded (Figure 2A), these data suggest that precursor-miRNA-stem loops are produced from both the sense (same orientation as XPA) and antisense strands. The stem loop precursor processed by Drosha from the sense RNA (tgu-mir-2954) generates two active miRNAs from its both arms (tgu-miR-2954-3p and tgu-miR-2954-5p). The stem loop precursor processed by Drosha from the antisense RNA (tgu-mir-2954R) generates at least one active miRNA (tgu-miR-2954R-5p) from its 5' end sequence.

We carried out Northern analysis on tgu-miR-2954-3p, which is the miRNA that has the highest number of read counts detected in our Illumina assays among the three miRNAs from the tgu-mir-2954 locus. A robust signal at ~22 nucleotides is evident in mixed-sex pools of RNA from birds hearing either song or silence, and a weaker signal is also detectable in two female-only pools of RNA (Figure 2C). Greater expression in males is consistent with the ZZ genotype of males and the lack of efficient sex chromosome dosage compensation in the zebra finch [37,38].

By TaqMan as well as by Illumina, we observed an apparent sex difference in the direction of the response of tgu-miR-2954-3p to song - up in males and down in females (Figure 3 and Additional File 1, Table S4). This suggests this locus may be under complex regulation, integrating information about sex, auditory or social experience and perhaps also other factors related to XPA gene expression.

To gain insight into the potential functional role of tgu-miR-2954-3p in the response to song, we used a conservative strategy to predict gene targets that are both conserved in birds and responsive to song exposure in the zebra finch. Potential targets of miRNAs are described as mRNAs that have sequences that can undergo Watson-Crick base pairing with the 5'-seed (nucleotide 2-7) of the miRNA [39]. For target prediction we applied the TargetScan (5.1) algorithm using the chicken genome as an initial reference, and then confirmed presence of the target sequence in the zebra finch. For evidence of song responsiveness, we used the data set of Dong et al. [14]. Eight genes met all these criteria (Table 3) and are thus both song-responsive and also subject to regulation by tgu-miR-2954-3p. These genes all have functions in control of cell proliferation or neurite outgrowth (see below).

Discussion

Here we show that a natural perceptual experience, hearing the sound of another bird singing, alters the

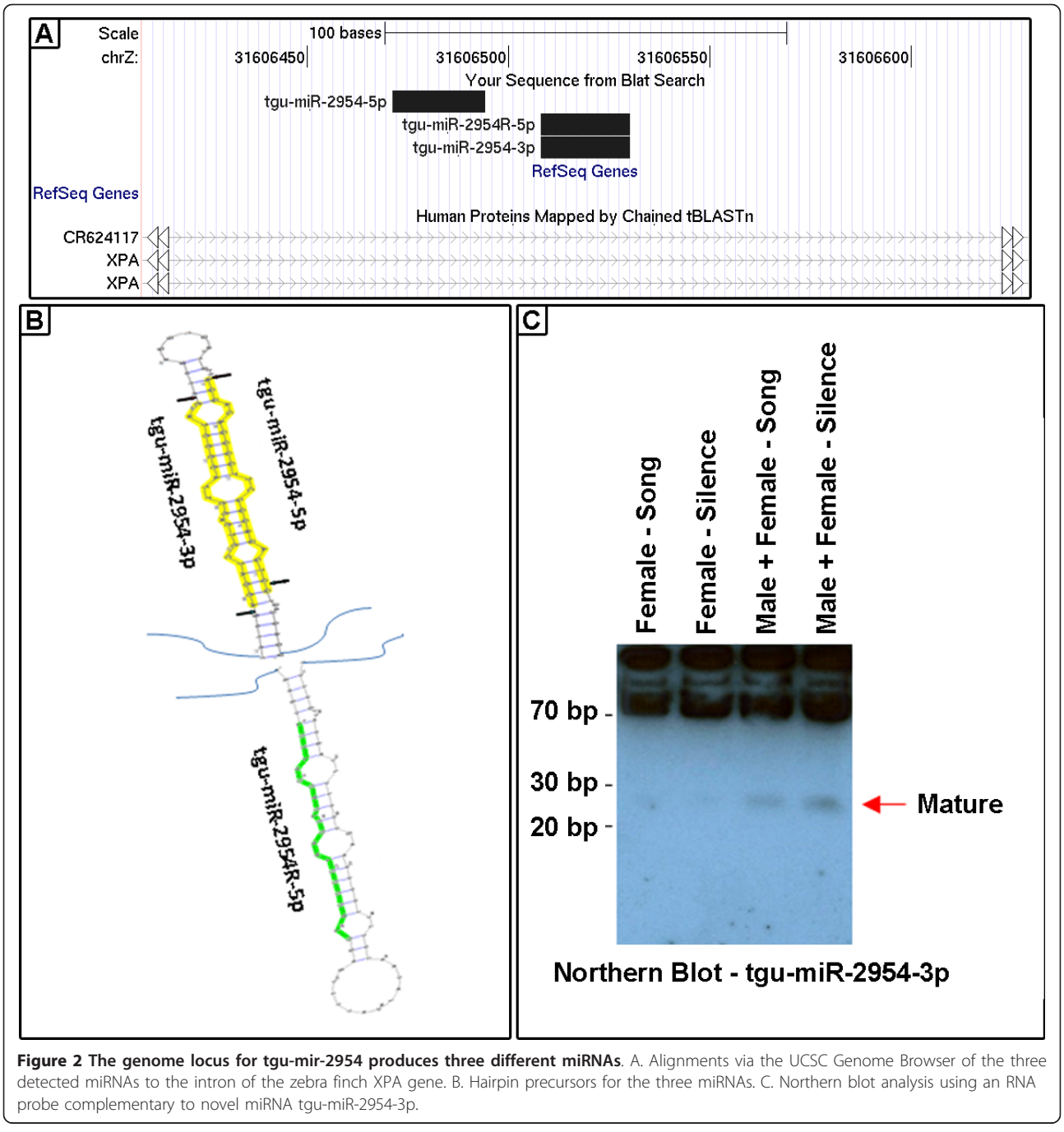
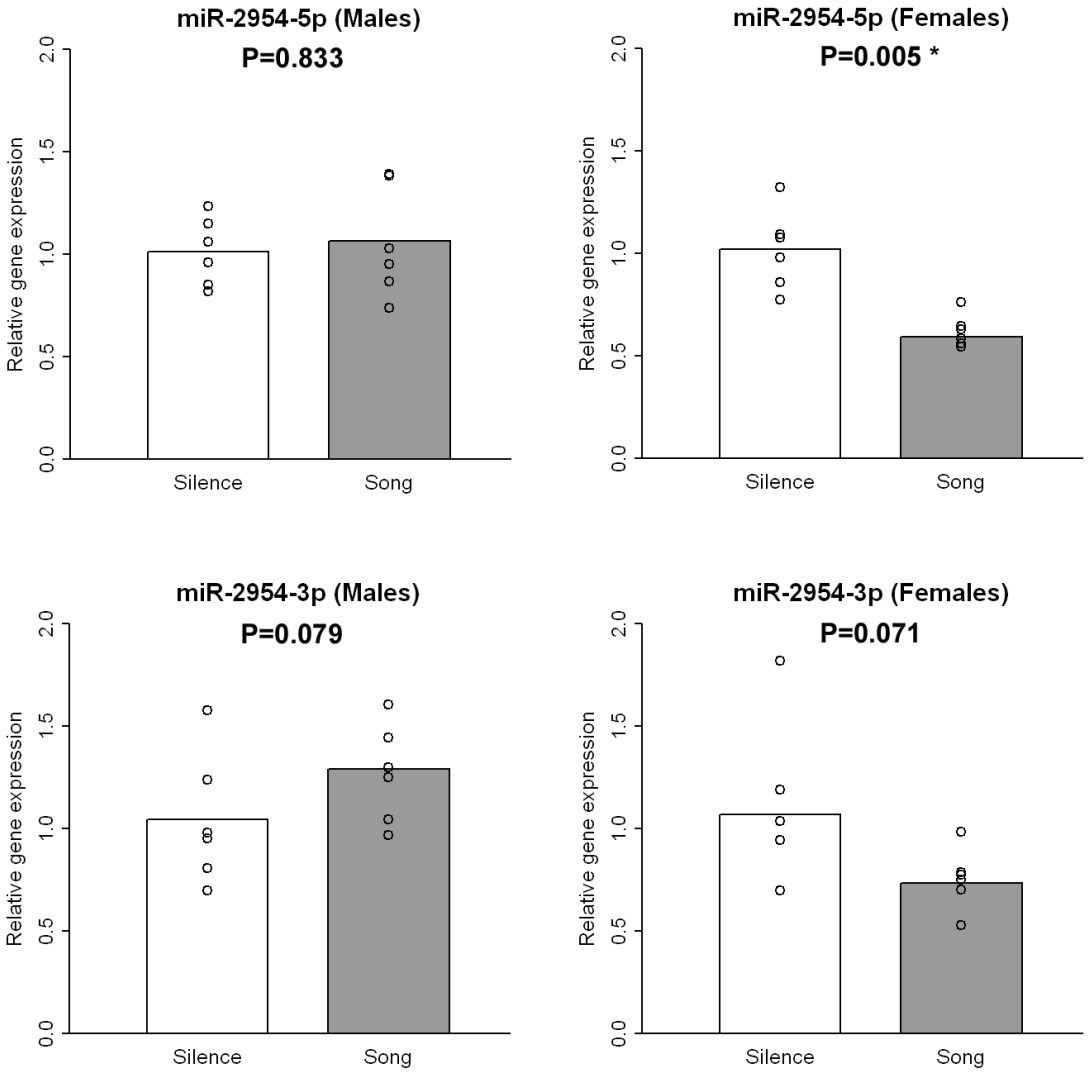


Figure 2 The genome locus for tgu-mir-2954 produces three different miRNAs. A. Alignments via the UCSC Genome Browser of the three detected miRNAs to the intron of the zebra finch XPA gene. B. Hairpin precursors for the three miRNAs. C. Northern blot analysis using an RNA probe complementary to novel miRNA tgu-miR-2954-3p.

profile of miRNAs in parts of the songbird brain responsible for auditory perception, integration and memory. The song-regulated population includes both known (conserved) and novel miRNAs. We highlight one sex-linked song-responsive miRNA and identify mRNAs that are potential targets of its action during song exposure. Thus miRNAs may have roles in the information processing functions of the brain, in addition to their roles in brain development and evolution.

To demonstrate this, we first catalogued the miRNAs expressed in the adult zebra finch auditory forebrain. We used massively parallel Illumina sequencing of small RNAs to perform this cataloguing efficiently. In addition to known conserved miRNAs, our analysis identified 48 small RNA sequences that meet the structural criteria for miRNAs but had not been described in miRBase in any organism at the time of our analysis. Fourteen of these are detected in the chicken or human genome

A. TaqMan



B. Illumina

	RNA-seq read counts			
	Male Silence	Male Song	Female Silence	Female Song
tgu-miR-2954-5p	0	5	204	88
tgu-miR-2954-3p	1372	1610	756	273

Figure 3 Analysis of miRNAs produced at the tgu-mir-2954 locus. TaqMan and Illumina RNA-seq data generated from independent sets of birds (n = 6 in each data set) for expression from the tgu-mir-2954 locus. A) TaqMan results, where the relative gene expression of each individual bird (open circle) was obtained by using the 2^-ddCt method [98]; the relative gene expression of either Silence (white bar) or Song (gray bar) group was the mean of six individuals; the P value was calculated by paired t test since each song stimulated animal was explicitly paired with a silence control animal collected simultaneously. B) Read counts from the Illumina RNA-seq for miR-2954-3p and miR-2954-5p (also shown in the Additional File 1, Table S4).

Table 3 Song-regulated targets of tgu-miR-2954-3p

Ensembl ID	Gene Symbol	EST	Gene Name
ENSTGUG00000001349	ELAVL2	CK313262	ELAV-like protein 2 (Hu-antigen B)(HuB)(ELAV-like neuronal protein 1)(Nervous system-specific RNA-binding protein Hel-N1)
ENSTGUG00000001404	LINGO2	DV957508	Leucine-rich repeat and immunoglobulin-like domain-containing nogo receptor-interacting protein 2 Precursor (Leucine-rich repeat neuronal protein 6C)(Leucine-rich repeat neuronal protein 3)
ENSTGUG00000003073	TLK2	CK305975	Serine/threonine-protein kinase tousled-like 2 (EC 2.7.11.1)(Tousled-like kinase 2)(PKU-alpha)
ENSTGUG00000008207	BTG1	CK303273	Protein BTG1 (B-cell translocation gene 1 protein)
ENSTGUG00000008540	CHD2	DV958991	Chromodomain-helicase-DNA-binding protein 2 (CHD-2)(EC 3.6.1.)(ATP-dependent helicase CHD2)
ENSTGUG00000010181	XP_002196848.1	CK304764	crk-like protein (v-crk avian sarcoma virus CT10 oncogene homolog-like) (CRKL)
ENSTGUG00000010364	NEGR1	DV954047	Neuronal growth regulator 1 Precursor
ENSTGUG00000011700	HMGB1	CK314519	High mobility group protein B1 (High mobility group protein 1)(HMG-1)

We used TargetScan to find binding sites of tgu-miR-2954-3p on eight chicken genes and here are listed the information of their homologous genes in the zebra finch genome including Ensembl IDs, Gene Symbols, EST (Accession numbers of song-regulated EST identified in the previous microarray study) and Gene Names (or aliases in parenthesis).

assemblies and may give rise to miRNAs that have not yet been described elsewhere due to low copy number, restricted tissue distribution or other factors. The remaining novel miRNAs, 34 in number, may be unique to the zebra finch or the songbird lineage. Few studies have attempted *de novo* identification of miRNAs from the brain [18] and ours is the first to report direct sequencing of songbird brain miRNAs. A previous study did identify precursor sequences for five conserved miRNAs in the developing zebra finch brain [40]. Also, in parallel with our own Illumina analysis, Li and her colleagues used 454 sequencing to identify miRNAs in the brain and liver of adult zebra finches. These different sets of annotations are compared and collated in a supplement to the analysis of the zebra finch genome assembly [33].

By comparing birds hearing novel song playbacks or silence, we found evidence for experience-dependent fluctuations in large numbers of miRNAs in the auditory forebrain. We performed three separate pairwise comparisons by Illumina, where all aspects of the experimental conditions were carefully counterbalanced between the two groups in each comparison. The three comparisons were not direct replications of each other, as each had a different sex ratio. Our reasons for varying the sex ratio were partly pragmatic (limited numbers of birds of the same sex that could be removed from our aviary) and partly analytical (males and females have different behavioral responses to songs). Some of the differences between the three sets of results may reflect real biological differences in the responses of males and females. Indeed, our Northern analysis of the tgu-miR-2954-3p confirms a sex difference in expression of this Z-linked miRNA gene. This is especially intriguing because we also obtained TaqMan evidence for both sense and antisense transcripts of this miRNA. One can imagine scenarios where different ratios of sense and antisense transcription occur in males (two copies of the gene) and females (one copy of the gene) with different consequences on the transcriptional networks affected by song exposure in the two sexes.

Ignoring the potential effects of sex, we identified five miRNAs that showed significant and consistent changes in response to song across all three Illumina comparisons. Three miRNAs consistently decreased after song (tgu-miR-92, tgu-miR-124, tgu-miR-129-5p) and two increased (tgu-miR-25, tgu-miR-192). The down-regulated miRNAs are at much higher abundance (> 1000 reads in each run) and perhaps for this reason we were more successful at detecting them and replicating their song regulation by TaqMan assay in subsequent experiments with additional groups of birds. The most abundant miRNA in our regulated set, tgu-miR-124, consistently met the statistical test for significant down-regulation by song, in each of six separate experiments (three Illumina comparisons, two TaqMan analyses in Additional File 1, Table S4, and the TaqMan comparison of song vs. SEN in Additional File 2, Figure 3).

In studies in other species, miR-124 has been linked to brain plasticity and development in several contexts. Chronic cocaine administration results in down-regulation of miR-124 in the rodent mesolimbic dopaminergic system [41]. In the developing chick neural tube, miR-124a is a component of a regulatory network that controls the transition between neural progenitors and post-mitotic neurons [42]. miR-124 also regulates adult neurogenesis, and its overexpression promotes neuronal differentiation [42,43] and neurite outgrowth [44]. Intriguingly, in songbirds neurogenesis continues in the forebrain throughout adulthood, from a population of precursor cells that line the walls of the lateral ventricles and have the characteristics of neural stem cells [45-47].

The net rate of neuronal addition and loss in the adult songbird has been shown to depend on social and environmental influences [48-51]. Perhaps tgu-miR-124 is a regulatory link between experience and neurogenesis - further study of this fascinating possibility is clearly warranted.

Although miRNAs can have diverse functions, they often act by altering the concentrations of specific mRNAs they target via complementary base pairing. We used the TargetScan algorithm [52] to predict binding sites of tgu-miR-2954-3p in chicken genes, and then we confirmed the presence of the same conserved target sequence in the zebra finch genome assembly. We found eight targets that met these criteria and were also regulated by song in the Dong et al. microarray data [14]. These eight genes have a provocative coherence in their function, as they are all implicated in control of cell proliferation and neuronal differentiation. Six operate by affecting gene expression and chromatin remodeling as we briefly review here. ELAVL2 is a member of a protein family that binds AU-rich regions in the 3'UTR of genes such as c-fos and promotes the shift from cell proliferation into cellular differentiation [53-57]. TLK2 is a kinase tightly associated with DNA replication during cell division [58]. At least one of its targets, the histone chaperone Asf1, controls chromatin assembly, thus TLK2 activity can regulate transcription and elongation [59-61]. BTG1 is also regulated during the cell cycle [62]. It acts as a cofactor for Hoxb9, a transcription factor that controls cell proliferation and differentiation, and BTG1 reduces rates of cell proliferation [62-64]. CHD2 can potentially affect transcription of many genes by remodeling chromatin [65,66]; disruption of CHD2 has profound consequences for development and is implicated in many human diseases [67-69]. HMGB1 is another DNA binding protein that facilitates transcription by altering chromatin structure to ease promoter binding [70-73]. Some of the genes regulated by HMGB1 may play a role in cell proliferation and migration [74,75]. Neuronal migration and neurite outgrowth are affected by CRKL, a transcriptional activator that is a component of the reelin pathway [76-79]. Unlike the other six genes, NEGR1 and LINGO2 do not seem to alter transcription but they do have established roles in neuronal differentiation. NEGR1 affects cell-cell adhesion to modulate neurite outgrowth and synapse formation [80-82]. LINGO2 is one member of a family of transmembrane proteins that are involved in neural and axonal regeneration [83,84]. The function of LINGO2 is untested, but expression of a related protein, LINGO1, is attenuated in cortical areas deprived of sensory input and is a partner in a signaling pathway that correlates with neuronal activity during a learning paradigm [85,86].

Conclusions

In conclusion, these data reveal a network of miRNAs in the zebra finch's auditory forebrain, responsive to the experience of hearing another bird sing. The network includes well-characterized conserved miRNA known to have roles in neuronal differentiation (miR-124), and novel miRNAs that can target genes that control neuronal differentiation (tgu-miR-2954-3p). Our data suggest this miRNA network may influence the fundamental shift we have observed in the transcriptional and metabolic state of the auditory forebrain during the process of song-specific habituation [14,87]. Further study of song responses in the zebra finch may reveal general insights into the neurogenomic mechanisms that underlie learning, memory and the ongoing adaptation to experience.

Methods

Song stimulation and brain dissections

Zebra finches were obtained from aviaries maintained at the University of Illinois. All procedures involving animals were conducted with the approval of the University of Illinois Institutional Animal Care and Use Committee. The birds were raised in a standard breeding aviary and were tutored under normal social conditions (i.e., by their parents or other adult birds in the breeding colony). All birds used in this study were adults (older than 90 days after hatching). The song playback procedures and brain dissections were performed exactly as in previous microarray analyses, using the same equipment [14,88]. Briefly, each bird was put individually into a sound isolation chamber for 18 hours on the first day, and on the second day those in the song group heard 30 minutes of a song not heard previously ("novel song"). Matched controls collected in parallel heard no song playback ("silence"). Birds were sacrificed in song-silence pairs, so that 5 minutes before the end of the song playback to one bird, a bird in the silence group was sacrificed and its auditory forebrain was dissected and frozen in dry ice. Then the auditory forebrain of the song-stimulated bird was dissected and frozen in dry ice. The auditory forebrain dissection (also referred to as auditory lobule) is described in [89] and collects NCM (caudomedial nidopallium), CMM (caudomedial mesopallium) and the enclosed Field L subregions. At the end of the song stimulation procedure, all auditory forebrains were transferred and stored at -80C until RNA isolation. For the comparison of responses after overnight isolation to song versus SEN (Additional File 2, Figure S3), we used two matched stimuli derived from bird "C7" as previously described [34].

RNA Samples

For Illumina analyses: Total RNA was extracted using the mirVana miRNA Isolation Kit (Ambion) from three

pairs of pooled auditory forebrain samples. 1) Males (samples S7 and S8): 6 birds per pool, collected in November 2008. 2) Females (samples S1 and S2): 6 birds per pool, collected in May 2009. 3) Mixed (samples S3 and S4), 3 males and 3 females each pool, collected in May 2009. Samples with odd numbers were from birds hearing song, and even number hearing silence.

For Northern analysis: Auditory forebrains of 22 birds (12 females and 10 males) were collected in April 2009, and total RNA was extracted by Tri-Reagent (Ambion). Male and female samples were pooled after extraction.

For TaqMan analysis: Analyses were performed on total RNA extracted either by mirVana or Tri-Reagent (Ambion), from the auditory forebrains of individual males or females, collected in April-August 2009, March 2010 or December 2010.

Illumina small RNA sequencing and novel miRNA discovery

Fifteen micrograms of total RNA from auditory forebrain of song bird samples described above were gel-fractionated to isolate 18-30 nt small RNAs. 3' and 5' adapters were ligated to the small RNAs and constructs amplified following RT-PCR following the conditions specified in the small RNA kit (FC-102-1009, Illumina) protocol. The small RNA library was sequenced using a Solexa/Illumina GA-1 Genome analyzer. Small RNA sequences were analyzed through a high-throughput computational pipeline described by [28,29,90,91]. To identify zebra finch miRNAs that are also conserved in chicken, human and mouse, we performed a local Smith-Waterman alignment of each unique sequence read against each of the mature miRNAs in miRBase version 13.0 for each of these species. We allowed for a 3 base overhang on the 5' end and a 6 base overhang on the 3' end. In the case of redundantly aligning reads, mature miRNA sequences were equally apportioned among each of the hairpins. For each sample, all sequence reads were aligned to a reference set of precursor miRNAs from miRBase version 13.0. The reads that did not align to any known miRNA were passed to our novel miRNA discovery platform as previously described [28]. Briefly, each sequence is first mapped to the reference genome sequence (WUGSC 3.2.4) and 200 bases of flanking sequence are extracted to further define the putative hairpin. This extracted sequence is then folded using the Vienna RNA folding package [92] and those sequences that form a plausible hairpin are selected as potential novel miRNA hairpins. These candidates are filtered through a set of three Ambros criteria: 1) the mature putative miRNA sequence must rest on one side of a single hairpin; 2) the putative miRNA sequence must bind relatively tightly within the hairpin stem containing no large or energetically unfavorable loops; and 3) the putative hairpin must have a miRNA-appropriate energy (free energy below -20 kcal/mol). All sequences that passed were then carefully curated to determine if Drosha and Dicer processing could yield the resulting mature sequence from the predicted hairpin. These candidates are then divided into four different categories: "not likely", "potential", "high confidence", and "confirmed" (as in red, gray, blue and green colors in Additional File 1, Tables S2 and S3). Candidates that are flagged red as "not likely" either failed to map in a pile of sequences in a very tight space of 15-25 nt of the predicted hairpin (e.g. were scattered evenly across the full length of the hairpin), mapped within the loop of the hairpin, or mapped to known tRNAs or rRNAs. Candidates that passed all of the above criteria, and also mapped within a hairpin with predicted Drosha and Dicer cut sites were categorized as "high confidence" (blue annotation in Additional File 1, Tables S2 and S3). All high confidence candidates for which we detected both the mature sequence and the putative star sequence from the same hairpin we categorized as "confirmed" (green annotation in Additional File 1, Table S3). In addition to miRNA precursors, the reads were also mapped to the reference zebra finch genome using the Pash software package [93,94], and uploaded to the Genboree platform (http://www.genboree.com) to identify potential mappings to piRNAs, snoRNAs and other annotations in addition to miRNAs (data shown in Additional File 1, Table S1). PiRNAs (i.e., Piwi-interacting RNAs) have a central role in the maintenance of the integrity of genomes through the silencing of transposable elements [95]. SnoRNAs (small nucleolar RNAs) function in site-specific ribosomal RNA modification, rRNA processing and more recently have been found to guide alternate splicing and RNA editing of mRNA transcripts [96].

TaqMan qPCR

To measure the mature miRNA, the TaqMan MicroRNA Assay Kit (Applied Biosystems) was used according to the manufacturer's instructions. Probe sequences used for each target miRNA are given in Table 4.

Northern Blot Analysis

Northern blotting to confirm novel miRNA tgu-miR-2954-3p was performed by modifying the protocol of [97]. 2 μg of total RNA was heated at 65°C for 5 min with 2X loading dye (Ambion), quenched on ice, and loaded on a 15% TBE Urea gel (Invitrogen). Total RNA was separated by electrophoresis at 200V for 50 min. The gel was stained with with EtBr in 1x TBE (4 μL of 10 mg/ml EtBr per 100 ml of 1x TBE) for 3 minutes with gentle shaking and transferred to nylon membrane for 90

Table 4 Probes used for Taqman analysis of specific miRNA sequences

miRBase name	Company name	Sequence detected
tgu-let-7a	let-7a	5'-UGAGGUAGUAGGUUGUAUAGUU-3'
tgu-let-7f	let-7f	5'-UGAGGUAGUAGAUUGUAUAGUU-3'
tgu-miR-124	miR-124	5'-UAAGGCACGCGGUGAAUGCC-3'
tgu-miR-9	miR-9	5'-UCUUUGGUUAUCUAGCUGUAUGA-3'
tgu-miR-129-5p	miR-129-5p	5'-CUUUUUGCGGUCUGGGCUUGC-3'
tgu-miR-129-3p	miR-129-3p	5'-AAGCCCUUACCCCAAAAAGCAU-3'
tgu-miR-29a	miR-29c	5'-UAGCACCAUUUGAAAUCGGU-3'
tgu-miR-92	miR-92a	5'-UAUUGCACUUGUCCCGGCCUGU-3'
tgu-miR-25	miR-25	5'-CAUUGCACUUGUCUCGGUCUGA-3'
RNU6B	RNU6B	5'-CGCAAGGAUGACACGCAAAUUCGUGAAGCGUUCCAUAUUUUU-3'
tgu-miR-2954-5p	novel51F-5p	5'-GCUGAGAGGGCUUGGGGAGAGGA-3'
tgu-miR-2954-3p	novel51F-3p	5'-CAUCCCCAUUCCACUCCUAGCA-3' (Northern validated)
tgu-miR-2954R-5p	novel51R-5p	5'-UGCUAGGAGUGGAAUGGGGAUG-3'
tgu-miR-2954R-3p	novel51R-3p	5'-UCCUCUCCCCAAGCCCUCUCAGC-3'

min at 200V using 1X TBE buffer at room temperature. The membrane was cross-linked at 1200 kJ for 45 seconds. RNA probes were synthesized for tgu-miR-2954-3p probe 5' - UGCUAGGAGUGGAAUGGGGAU G - 3' by Integrated DNA Technologies. Radio labeling was carried out in a reaction of 12.0ul dH2O + 2.0ul PNK buffer + 1.0ul (100ng/ul) probe + 1.0ul PNK polymerase (Promega) + 4.0ul P^{32}-gamma-ATP (10mCi/ml) (PerkinElmer). The reaction was incubated at 37°C for 1 hour and inactivated at 65°C for 10 min. The probe was purified using Nick columns from GE following manufacturer's instructions. The membranes were pre-hybridized for 30 min with 20 ml of pre-hybridization buffer (5X SSC + 20 mM NaPO4 + 7X SDS + 2X Denhardt (pre warmed) at 60° C) in a rotating hybridization oven. Hybridization was carried out at 50°C in a rotating incubator for 24h. The membranes were washed for 10 min at 50°C with 20-30mL of wash buffer (2X SSC + 0.5% SDS). When background was ~0.5 cpm, the membranes were wrapped in saran wrap and exposed at -80°C for ~72h.

Additional material

Additional file 1: Supplemental tables.xls. This one file contains all four Supplemental Tables, each as a separate worksheet. **Table S1 ("1 overview")** is a summary of Illumina sequence read alignments for six pools of RNA from zebra finch auditory forebrain responding to song versus silence, and shows the distribution of sequence reads in relation to multiple genomes and multiple annotations in the current genomic databases. **Table S2 ("2 novel hairpins")** gives detailed alignments of putative pre-miRNAs and read sequences. **Table S3 ("3 novel genes")** shows annotations of all novel miRNA loci mapped in genome assemblies of zebra finch, chicken or human. **Table S4 ("4 all read counts")** gives read counts and current annotation in miRBase of all conserved and novel miRNAs, with statistics.

Additional file 2: Supplemental figures.doc. This one file contains all three supplemental figures. **Figure S1** is a Venn diagram of numbers of miRNAs with significant differential expression in response to novel song in three Illumina experiments. **Figure S2** shows a comparative mapping in other avian transcriptomes of tgu-mir-2954. **Figure S3** demonstrates the song-specificity of the miRNA response, using TaqMan to compare the levels of specific miRNAs in animals from groups that heard song, matching song-enveloped noise, or silence.

Acknowledgements
We thank Sarah London for useful discussions and contributions to the text. Supported by NIH RO1 NS045264 and RO1 NS051820 (to D.F.C).

Note Added in Proof:
The novel miRNA referred to here as "miR-2954-3p" is now identified in miRBase as "miR-2954". The novel miRNA referred to here as "miR-2954-5p" is now identified in miRBase as "miR-2954*".

Author details
[1]Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204, USA. [2]Departments of Pathology, Baylor College of Medicine, Houston, Texas 77030, USA. [3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [4]Department of Cell and Developmental Biology, University of Illinois, Urbana-Champaign, IL 61801, USA. [5]W.M. Keck Center for Comparative and Functional Genomics, Roy J. Carver Biotechnology Center, University of Illinois, Urbana-Champaign, IL 61801, USA. [6]Dan Duncan Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA. [7]ARK-Genomics, The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, EH25 9RG, UK. [8]Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK. [9]Institute for Genomic Biology, University of Illinois, Urbana-Champaign, IL 61801, USA. [10]Beckman Institute, University of Illinois, Urbana-Champaign, IL 61801, USA. [11]Bioinformatics Research Laboratory (BRL), Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA.

Authors' contributions
PHG coordinated the work of Illumina RNA-seq and prepared the manuscript. YL conducted the song exposure experiments and subsequent dissections and RNA extractions, performed TaqMan qPCR, analyzed differentially expressed miRNAs and participated in drafting the manuscript. ALB performed Illumina RNA-seq and Northern blot. JD helped analyze expression data of Illumina RNA-seq and TaqMan qPCR. CC, JBT, CJC, JHK, and AM participated in mapping and analyzing Illumina RNA-seq data. MW and SGJ helped with miRNA sequence annotation. DFC designed and coordinated the study and drafted the manuscript. All authors read and approved the manuscript.

Received: 28 July 2010 Accepted: 31 May 2011 Published: 31 May 2011

References

1. Clayton DF, Balakrishnan CN, London SE: **Integrating genomes, brain and behavior in the study of songbirds.** *Curr Biol* 2009, **19(18)**:R865-873.
2. Jarvis ED: **Learned birdsong and the neurobiology of human language.** *Ann N Y Acad Sci* 2004, **1016**:749-777.
3. Miller DB: **Acoustic Basis of Mate Recognition by Female Zebra Finches (Taeniopygia-Guttata).** *Animal Behaviour* 1979, **27(May)**:376-380.
4. Miller DB: **Long-Term Recognition of Fathers Song by Female Zebra Finches.** *Nature* 1979, **280(5721)**:389-391.
5. Clayton NS: **Song Discrimination-Learning in Zebra Finches.** *Animal Behaviour* 1988, **36**:1016-1024.
6. Stripling R, Kruse AA, Clayton DF: **Development of song responses in the zebra finch caudomedial neostriatum: Role of genomic and electrophysiological activities.** *Journal of Neurobiology* 2001, **48(3)**:163-180.
7. Mello CV, Vicario DS, Clayton DF: **Song presentation induces gene expression in the songbird forebrain.** *Proc Natl Acad Sci USA* 1992, **89(15)**:6818-6822.
8. Mello C, Nottebohm F, Clayton D: **Repeated exposure to one song leads to a rapid and persistent decline in an immediate early gene's response to that song in zebra finch telencephalon.** *J Neurosci* 1995, **15(10)**:6919-6925.
9. Jarvis ED, Scharff C, Grossman MR, Ramos JA, Nottebohm F: **For whom the bird sings: context-dependent gene expression.** *Neuron* 1998, **21(4)**:775-788.
10. Clayton DF: **The genomic action potential.** *Neurobiol Learn Mem* 2000, **74(3)**:185-216.
11. Kruse AA, Stripling R, Clayton DF: **Context-specific habituation of the zenk gene response to song in adult zebra finches.** *Neurobiol Learn Mem* 2004, **82(2)**:99-108.
12. Vignal C, Andru J, Mathevon N: **Social context modulates behavioural and brain immediate early gene responses to sound in male songbird.** *Eur J Neurosci* 2005, **22(4)**:949-955.
13. Woolley SC, Doupe AJ: **Social context-induced song variation affects female behavior and gene expression.** *PLoS Biol* 2008, **6(3)**:e62.
14. Dong S, Replogle KL, Hasadsri L, Imai BS, Yau PM, Rodriguez-Zas S, Southey BR, Sweedler JV, Clayton DF: **Discrete molecular states in the brain accompany changing responses to a vocal signal.** *Proc Natl Acad Sci USA* 2009, **106(27)**:11364-11369.
15. London SE, Dong S, Replogle K, Clayton DF: **Developmental shifts in gene expression in the auditory forebrain during the sensitive period for song learning.** *Dev Neurobiol* 2009, **69(7)**:437-450.
16. Cao X, Yeo G, Muotri AR, Kuwabara T, Gage FH: **Noncoding RNAs in the mammalian central nervous system.** *Annu Rev Neurosci* 2006, **29**:77-103.
17. Ason B, Darnell DK, Wittbrodt B, Berezikov E, Kloosterman WP, Wittbrodt J, Antin PB, Plasterk RH: **Differences in vertebrate microRNA expression.** *Proc Natl Acad Sci USA* 2006, **103(39)**:14385-14389.
18. Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RH: **Diversity of microRNAs in human and chimpanzee brain.** *Nat Genet* 2006, **38(12)**:1375-1377.
19. Bak M, Silahtaroglu A, Moller M, Christensen M, Rath MF, Skryabin B, Tommerup N, Kauppinen S: **MicroRNA expression in the adult mouse central nervous system.** *RNA* 2008, **14(3)**:432-444.
20. Krichevsky AM, King KS, Donahue CP, Khrapko K, Kosik KS: **A microRNA array reveals extensive regulation of microRNAs during brain development.** *RNA* 2003, **9(10)**:1274-1281.
21. Miska EA, Alvarez-Saavedra E, Townsend M, Yoshii A, Sestan N, Rakic P, Constantine-Paton M, Horvitz HR: **Microarray analysis of microRNA expression in the developing mammalian brain.** *Genome Biol* 2004, **5(9)**: R68.
22. Sempere LF, Freemantle S, Pitha-Rowe I, Moss E, Dmitrovsky E, Ambros V: **Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation.** *Genome Biol* 2004, **5(3)**:R13.
23. Li N, Bates DJ, An J, Terry DA, Wang E: **Up-regulation of key microRNAs, and inverse down-regulation of their predicted oxidative phosphorylation target genes, during aging in mouse brain.** *Neurobiol Aging* 2009.
24. Schratt G: **Fine-tuning neural gene expression with microRNAs.** *Curr Opin Neurobiol* 2009, **19(2)**:213-219.
25. Cheng HY, Papp JW, Varlamova O, Dziema H, Russell B, Curfman JP, Nakazawa T, Shimizu K, Okamura H, Impey S, *et al*: **microRNA modulation of circadian-clock period and entrainment.** *Neuron* 2007, **54(5)**:813-829.
26. Schratt GM, Tuebing F, Nigh EA, Kane CG, Sabatini ME, Kiebler M, Greenberg ME: **A brain-specific microRNA regulates dendritic spine development.** *Nature* 2006, **439(7074)**:283-289.
27. Fiore R, Khudayberdiev S, Christensen M, Siegel G, Flavell SW, Kim TK, Greenberg ME, Schratt G: **Mef2-mediated transcription of the miR379-410 cluster regulates activity-dependent dendritogenesis by fine-tuning Pumilio2 protein levels.** *EMBO J* 2009, **28(6)**:697-710.
28. Creighton CJ, Reid JG, Gunaratne PH: **Expression profiling of microRNAs by deep sequencing.** *Brief Bioinform* 2009, **10(5)**:490-497.
29. Reid JG, Nagaraja AK, Lynn FC, Drabek RB, Muzny DM, Shaw CA, Weiss MK, Naghavi AO, Khan M, Zhu H, *et al*: **Mouse let-7 miRNA populations exhibit RNA editing that is constrained in the 5'-seed/cleavage/anchor regions and stabilize predicted mmu-let-7a:mRNA duplexes.** *Genome Res* 2008, **18(10)**:1571-1581.
30. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, *et al*: **Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.** *Genome Res* 2008, **18(4)**:610-621.
31. Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, *et al*: **A mammalian microRNA expression atlas based on small RNA library sequencing.** *Cell* 2007, **129(7)**:1401-1414.
32. Lee RC, Ambros V: **An extensive class of small RNAs in Caenorhabditis elegans.** *Science* 2001, **294(5543)**:862-864.
33. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, *et al*: **The genome of a songbird.** *Nature* 2010, **464(7289)**:757-762.
34. Park KH, Clayton DF: **Influence of restraint and acute isolation on the selectivity of the adult zebra finch zenk gene response to acoustic stimuli.** *Behav Brain Res* 2002, **136(1)**:185-191.
35. Bailey D, Wade J: **Differential expression of the immediate early genes FOS and ZENK following auditory stimulation in the juvenile male and female zebra finch.** *Brain Res Mol Brain Res* 2003, **116(1-2)**:147-154.
36. K√ºnstner A, Wolf JBW, Backstr√∂m N, Whitney O, Balakrishnan CN, Day L, Edwards SV, Janes DE, Schlinger BA, Wilson RK, *et al*: **Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species.** *Molecular Ecology* 2010, **19(SUPPL. 1)**:266-276.
37. Itoh Y, Melamed E, Yang X, Kampf K, Wang S, Yehya N, Van Nas A, Replogle K, Band MR, Clayton DF, *et al*: **Dosage compensation is less effective in birds than in mammals.** *J Biol* 2007, **6(1)**:2.
38. Itoh Y, Replogle K, Kim YH, Wade J, Clayton DF, Arnold AP: **Sex bias and dosage compensation in the zebra finch versus chicken genomes: General and specialized patterns among birds.** *Genome Research* 2010, **20(4)**:512-518.
39. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell* 2009, **136(2)**:215-233.
40. Li X, Wang XJ, Tannenhauser J, Podell S, Mukherjee P, Hertel M, Biane J, Masuda S, Nottebohm F, Gaasterland T: **Genomic resources for songbird research and their use in characterizing gene expression during brain development.** *Proc Natl Acad Sci USA* 2007, **104(16)**:6834-6839.
41. Chandrasekar V, Dreyer JL: **microRNAs miR-124, let-7d and miR-181a regulate cocaine-induced plasticity.** *Mol Cell Neurosci* 2009, **42(4)**:350-362.
42. Visvanathan J, Lee S, Lee B, Lee JW, Lee SK: **The microRNA miR-124 antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development.** *Genes Dev* 2007, **21(7)**:744-749.
43. Cheng LC, Pastrana E, Tavazoie M, Doetsch F: **miR-124 regulates adult neurogenesis in the subventricular zone stem cell niche.** *Nat Neurosci* 2009, **12(4)**:399-408.
44. Yu JY, Chung KH, Deo M, Thompson RC, Turner DL: **MicroRNA miR-124 regulates neurite outgrowth during neuronal differentiation.** *Exp Cell Res* 2008, **314(14)**:2618-2633.
45. Goldman SA, Nottebohm F: **Neuronal production, migration and differentiation in a vocal control nucleus of the adult female canary brain.** *Proceedings of the National Academy of Sciences of the United States of America* 1983, **80**:2390-2394.

46. Alvarez-Buylla A, Theelen M, Nottebohm F: **Proliferation "hot spots" in adult avian ventricular zone reveal radial cell division.** *Neuron* 1990, **5(1)**:101-109.
47. Alvarez-Buylla A, Kirn JR: **Birth, migration, incorporation, and death of vocal control neurons in adult songbirds.** *Journal of Neurobiology* 1997, **33**:585-601.
48. Barnea A: **Interactions between environmental changes and brain plasticity in birds.** *General and Comparative Endocrinology* 2009, **163(1-2)**:128-134.
49. Kirn JR: **The relationship of neurogenesis and growth of brain regions to song learning.** *Brain and Language* 2010, **115(1)**:29-44.
50. Nottebohm F, O'Loughlin B, Gould K, Yohay K, Alvarez-Buylla A: **The life span of new neurons in a song control nucleus of the adult canary brain depends on time of year when these cells are born.** *Proceedings of the National Academy of Sciences of the United States of America* 1994, **91(17)**:7849-7853.
51. Wilbrecht L, Crionas A, Nottebohm F: **Experience affects recruitment of new neurons but not adult neuron number.** *Journal of Neuroscience* 2002, **22(3)**:825-831.
52. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120(1)**:15-20.
53. Levine TD, Gao F, King PH, Andrews LG, Keene JD: **Hel-N1: an autoimmune RNA-binding protein with specificity for 3' uridylate-rich untranslated regions of growth factor mRNAs.** *Mol Cell Biol* 1993, **13(6)**:3494-3504.
54. Abe R, Yamamoto K, Sakamoto H: **Target specificity of neuronal RNA-binding protein, Mel-N1: direct binding to the 3' untranslated region of its own mRNA.** *Nucleic Acids Res* 1996, **24(11)**:2011-2016.
55. Ma WJ, Chung S, Furneaux H: **The Elav-like proteins bind to AU-rich elements and to the poly(A) tail of mRNA.** *Nucleic Acids Res* 1997, **25(18)**:3564-3569.
56. Akamatsu W, Okano HJ, Osumi N, Inoue T, Nakamura S, Sakakibara S, Miura M, Matsuo N, Darnell RB, Okano H: **Mammalian ELAV-like neuronal RNA-binding proteins HuB and HuC promote neuronal development in both the central and the peripheral nervous systems.** *Proc Natl Acad Sci USA* 1999, **96(17)**:9885-9890.
57. Hambardzumyan D, Sergent-Tanguy S, Thinard R, Bonnamain V, Masip M, Fabre A, Boudin H, Neveu I, Naveilhan P: **AUF1 and Hu proteins in the developing rat brain: implication in the proliferation and differentiation of neural progenitors.** *J Neurosci Res* 2009, **87(6)**:1296-1309.
58. Sillje HH, Takahashi K, Tanaka K, Van Houwe G, Nigg EA: **Mammalian homologues of the plant Tousled gene code for cell-cycle-regulated kinases with maximal activities linked to ongoing DNA replication.** *EMBO J* 1999, **18(20)**:5691-5702.
59. Sillje HH, Nigg EA: **Identification of human Asf1 chromatin assembly factors as substrates of Tousled-like kinases.** *Curr Biol* 2001, **11(13)**:1068-1073.
60. Blackwell TK, Walker AK: **Transcription elongation: TLKing to chromatin?** *Curr Biol* 2003, **13(23)**:R915-916.
61. Carrera P, Moshkin YM, Gronke S, Sillje HH, Nigg EA, Jackle H, Karch F: **Tousled-like kinase functions with the chromatin assembly pathway regulating nuclear divisions.** *Genes Dev* 2003, **17(20)**:2578-2590.
62. Rouault JP, Puisieux A, Samarut C, Guehenneux F, Berthet C, Rimokh R, Falette N, Magaud JP: **Involvement of the BTG genes family in the control of cell cycle and DNA repair.** *Experimental Hematology* 1997, **25(8)**:229-229.
63. Corjay MH, Kearney MA, Munzer DA, Diamond SM, Stoltenborg JK: **Antiproliferative gene BTG1 is highly expressed in apoptotic cells in macrophage-rich areas of advanced lesions in Watanabe heritable hyperlipidemic rabbit and human.** *Laboratory Investigation* 1998, **78(7)**:847-858.
64. Li F, Liu J, Park ES, Jo M, Curry TE Jr: **The B cell translocation gene (BTG) family in the rat ovary: hormonal induction, regulation, and impact on cell cycle kinetics.** *Endocrinology* 2009, **150(8)**:3894-3902.
65. Hall JA, Georgel PT: **CHD proteins: a diverse family with strong ties.** *Biochem Cell Biol* 2007, **85(4)**:463-476.
66. Marfella CG, Imbalzano AN: **The Chd family of chromatin remodelers.** *Mutat Res* 2007, **618(1-2)**:30-40.
67. Bandres E, Malumbres R, Cubedo E, Honorato B, Zarate R, Labarga A, Gabisu U, Sola JJ, Garcia-Foncillas J: **A gene signature of 8 genes could identify the risk of recurrence and progression in Dukes' B colon cancer patients.** *Oncology Reports* 2007, **17(5)**:1089-1094.
68. Kulkarni S, Nagarajan P, Wall J, Donovan DJ, Donell RL, Ligon AH, Venkatachalam S, Quade BJ: **Disruption of chromodomain helicase DNA binding protein 2 (CHD2) causes scoliosis.** *Am J Med Genet A* 2008, **146A(9)**:1117-1127.
69. Nagarajan P, Onami TM, Rajagopalan S, Kania S, Donnell R, Venkatachalam S: **Role of chromodomain helicase DNA-binding protein 2 in DNA damage response signaling and tumorigenesis.** *Oncogene* 2009, **28(8)**:1053-1062.
70. Bustin M, Reeves R: **High-mobility-group chromosomal proteins: Architectural components that facilitate chromatin function.** *Progress in Nucleic Acid Research and Molecular Biology* 1996, **54**:35-100, Vol 54.
71. Grasser KD: **HMG1 and HU proteins: architectural elements in plant chromatin.** *Trends in Plant Science* 1998, **3(7)**:260-265.
72. Hall J, Thomas KL, Everitt BJ: **Cellular imaging of zif268 expression in the hippocampus and amygdala during contextual and cued fear memory retrieval: Selective activation of hippocampal CA1 neurons during the recall of contextual memories.** *Journal of Neuroscience* 2001, **21(6)**:2186-2193.
73. Bustin M: **At the crossroads of necrosis and apoptosis: signaling to multiple cellular targets by HMGB1.** *Sci STKE* 2002, **2002(151)**:pe39.
74. Guazzi S, Strangio A, Franzi AT, Bianchi ME: **HMGB1, an architectural chromatin protein and extracellular signalling factor, has a spatially and temporally restricted expression pattern in mouse brain.** *Gene Expression Patterns* 2003, **3(1)**:29-33.
75. Bassi R, Giussani P, Anelli V, Colleoni T, Pedrazzi M, Patrone M, Viani P, Sparatore B, Melloni E, Riboni L: **HMGB1 as an autocrine stimulus in human T98G glioblastoma cells: role in cell growth and migration.** *Journal of Neuro-Oncology* 2008, **87(1)**:23-33.
76. Ballif BA, Arnaud L, Arthur WT, Guris D, Imamoto A, Cooper JA: **Activation of a Dab1/CrkL/C3G/Rap1 pathway in Reelin-stimulated neurons.** *Curr Biol* 2004, **14(7)**:606-610.
77. Yip YP, Kronstadt-O'Brien P, Capriotti C, Cooper JA, Yip JW: **Migration of sympathetic preganglionic neurons in the spinal cord is regulated by reelin-dependent Dab1 tyrosine phosphorylation and CrkL.** *Journal of Comparative Neurology* 2007, **502(4)**:635-643.
78. Matsuki T, Pramatarova A, Howell BW: **Reduction of Crk and CrkL expression blocks reelin-induced dendritogenesis.** *J Cell Sci* 2008, **121(Pt 11)**:1869-1875.
79. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, *et al*: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37(Database)**:D690-697.
80. Brennan PA, Schellinck HM, Keverne EB: **Patterns of expression of the immediate-early gene egr-1 in the accessory olfactory bulb of female mice exposed to pheromonal constituents of male urine.** *Neuroscience* 1999, **90(4)**:1463-1470.
81. Schafer M, Brauer AU, Savaskan NE, Rathjen FG, Brummendorf T: **Neurotractin/kilon promotes neurite outgrowth and is expressed on reactive astrocytes after entorhinal cortex lesion.** *Molecular and Cellular Neuroscience* 2005, **29(4)**:580-590.
82. Hashimoto T, Yamada M, Maekawa S, Nakashima T, Miyata S: **IgLON cell adhesion molecule Kilon is a crucial modulator for synapse number in hippocampal neurons.** *Brain Research* 2008, **1224**:1-11.
83. Ishii N, Wanaka A, Tohyama M: **Increased expression of NLRR-3 mRNA after cortical brain injury in mouse.** *Brain Res Mol Brain Res* 1996, **40(1)**:148-152.
84. Bormann P, Roth LWA, Andel D, Ackermann M, Reinhard E: **zfNLRR, a novel leucine-rich repeat protein is preferentially expressed during regeneration in zebrafish.** *Molecular and Cellular Neuroscience* 1999, **13(3)**:167-179.
85. Josephson A, Trifunovski A, Scheele C, Widenfalk J, Wahlestedt C, Brene S, Olson L, Spenger C: **Activity-induced and developmental downregulation of the Nogo receptor.** *Cell Tissue Res* 2003, **311(3)**:333-342.
86. Endo T, Spenger C, Tominaga T, Brene S, Olson L: **Cortical sensory map rearrangement after spinal cord injury: fMRI responses linked to Nogo signalling.** *Brain* 2007, **130(Pt 11)**:2951-2961.
87. Dong S, Clayton DF: **Habituation in songbirds.** *Neurobiol Learn Mem* 2009, **92(2)**:183-188.
88. Replogle K, Arnold AP, Ball GF, Band M, Bensch S, Brenowitz EA, Dong S, Drnevich J, Ferris M, George JM, *et al*: **The Songbird Neurogenomics**

(SoNG) Initiative: community-based tools and strategies for study of brain gene function and evolution. *BMC Genomics* 2008, **9**:131.
89. Cheng HY, Clayton DF: **Activation and habituation of extracellular signal-regulated kinase phosphorylation in zebra finch auditory forebrain during song presentation.** *Journal of Neuroscience* 2004, **24(34)**:7503-7513.
90. Nagaraja AK, Andreu-Vieyra C, Franco HL, Ma L, Chen R, Han DY, Zhu H, Agno JE, Gunaratne PH, DeMayo FJ, *et al*: **Deletion of Dicer in somatic cells of the female reproductive tract causes sterility.** *Mol Endocrinol* 2008, **22(10)**:2336-2352.
91. Ma L, Buchold GM, Greenbaum MP, Roy A, Burns KH, Zhu H, Han DY, Harris RA, Coarfa C, Gunaratne PH, *et al*: **Correction: GASZ Is Essential for Male Meiosis and Suppression of Retrotransposon Expression in the Male Germline.** *PLoS Genet* 2009, **5(12)**.
92. Schuster P, Fontana W, Stadler PF, Hofacker IL: **From sequences to shapes and back: a case study in RNA secondary structures.** *Proc Biol Sci* 1994, **255(1344)**:279-284.
93. Kalafus KJ, Jackson AR, Milosavljevic A: **Pash: efficient genome-scale sequence anchoring by Positional Hashing.** *Genome Res* 2004, **14(4)**:672-678.
94. Coarfa C, Milosavljevic A: **Pash 2.0: scaleable sequence anchoring for next-generation sequencing technologies.** *Pac Symp Biocomput* 2008, 102-113.
95. Thomson T, Lin H: **The biogenesis and function of PIWI proteins and piRNAs: progress and prospect.** *Annu Rev Cell Dev Biol* 2009, **25**:355-376.
96. Royo H, Cavaille J: **Non-coding RNAs in imprinted gene clusters.** *Biol Cell* 2008, **100(3)**:149-166.
97. Gu P, Reid JG, Gao X, Shaw CA, Creighton C, Tran PL, Zhou X, Drabek RB, Steffen DL, Hoang DM, *et al*: **Novel microRNA candidates and miRNA-mRNA pairs in embryonic stem (ES) cells.** *PLoS One* 2008, **3(7)**:e2548.
98. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.** *Methods* 2001, **25(4)**:402-408.

doi:10.1186/1471-2164-12-277
Cite this article as: Gunaratne *et al.*: Song exposure regulates known and novel microRNAs in the zebra finch auditory forebrain. *BMC Genomics* 2011 **12**:277.

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 15 2013, pages 1902–1903
doi:10.1093/bioinformatics/btt297

Sequence analysis

Advance Access publication May 24, 2013

viRome: an R package for the visualization and analysis of viral small RNA sequence datasets

Mick Watson[1,*], Esther Schnettler[2] and Alain Kohl[2]

[1]ARK-Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG and [2]MRC-University of Glasgow Centre for Virus Research, 8 Church Street, Glasgow G11 5JR, UK

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: RNA interference (RNAi) is known to play an important part in defence against viruses in a range of species. Second-generation sequencing technologies allow us to assay these systems and the small RNAs that play a key role with unprecedented depth. However, scientists need access to tools that can condense, analyse and display the resulting data. Here, we present viRome, a package for R that takes aligned sequence data and produces a range of essential plots and reports.

Availability and implementation: viRome is released under the BSD license as a package for R available for both Windows and Linux http://virome.sf.net. Additional information and a tutorial is available on the ARK-Genomics website: http://www.ark-genomics.org/bioinformatics/virome.

Contact: mick.watson@roslin.ed.ac.uk

Received on December 20, 2012; revised on May 17, 2013; accepted on May 21, 2013

1 INTRODUCTION

RNA interference (RNAi) is mediated by small RNAs, such as micoRNAs (miRNAs) of 21–22 nt (Lagos-Quintana *et al.*, 2001), small interfering RNAs (siRNAs) of 21–22 nt (Bernstein *et al.*, 2001; Zamore *et al.*, 2000) and PIWI-interacting RNAs (piRNAs) of 24–30 nt (Aravin *et al.*, 2003; Brennecke *et al.*, 2007), and these molecules regulate many biological processes. These pathways are also a major part of the antiviral response in both insects and plants, including a variety of important mosquito-borne diseases of humans and animals, such as West Nile Virus, Dengue Virus and Chikungunya Virus. In arthropods, these are characterized by the production of 21–22 nt virus-derived small interfering RNAs (viRNAs) or 24–30 nt viral piRNA-like molecules (Blair, 2011; Donald *et al.*, 2012; Myles *et al.*, 2009).

Second-generation sequencing allows scientists to assay these systems in unprecedented depth, and short reads capture both the 21–22 nt siRNAs and the 24–30 nt piRNAs. However, there is a need for scientists to be able to summarize, analyse and visualize the results of such experiments. Here, we present viRome, a package for R, which takes aligned sequencing data in the BAM format (Li *et al.*, 2009) and produces a variety of plots and reports that are essential to the analysis of data from viral siRNA datasets.

*To whom correspondence should be addressed.

Software packages to analyse viral siRNA data exist. Paparrazi (Vodovar *et al.*, 2011) is designed to reconstruct viral genomes from siRNA data and produces some similar plots to viRome. Alternatively, Visitor (Antoniewski, 2011), an informatic pipeline for analysing short-read viRNA data, also produces several similar plots. However, both are implemented in Perl and are limited to the Linux/Unix operating system; they include alignment as part of the analysis; therefore, using an alternative aligner would require programming skills; finally, the plots are generated in batch mode; hence, there is no interaction between the user and the software.

As a package for R, viRome improves on these software packages in several ways, including (i) viRome allows interaction between the user and the software during report and graph generation, (ii) viRome is available on any operating system that supports R and has been tested on Microsoft Windows and several Linux distributions, (iii) viRome separates visualization from alignment; therefore, the user is free to use any alignment software they wish and (iv) as an R package, viRome integrates seamlessly with other R packages from the Bioconductor project (Gentleman *et al.*, 2004).

2 ANALYSIS AND VISUALIZATION

As input, viRome takes aligned sequence data in the BAM format. Many tools exist for alignment (Fonseca *et al.*, 2012) and provided they support the SAM/BAM format, viRome is capable of working with their output. Many of the functions within viRome attempt to summarize millions of data points into tables and plots that allow biological interpretation. One of the benefits of viRome is that most functions return the summarized data, as well as creating a plot. This allows users to create their own plots if they wish. Figure 1 shows a selection of plots produced by viRome.

Global analyses: One of the first requirements is to plot a histogram of the lengths of mapped reads—a peak at 21–22 nt implying an siRNA response, and a high frequency of 24–30 nt with a peak at 28 a piRNA response. In viRome, this can be created using the *barplot.bam* function. Users may also create a report using the *sequence.report* function. This produces a data.frame in R that summarizes and counts the sequences aligned to each base in a given reference sequence. Users can see the exact sequence, its length, the location and strand of the alignment plus a count of how many times that sequence

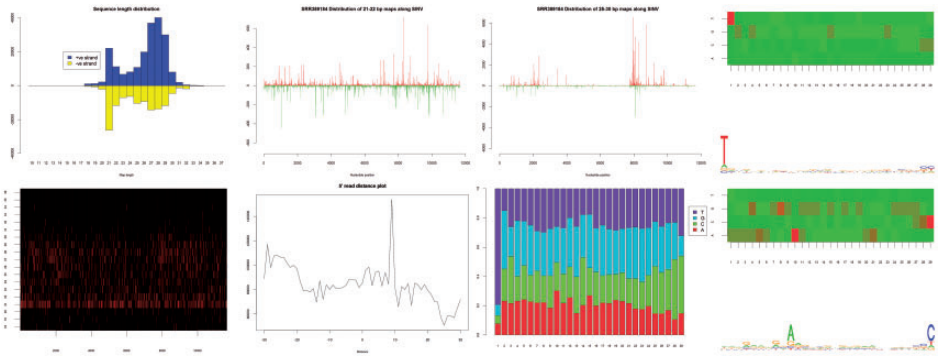


Fig. 1. Clockwise from top-left: a plot of read-length distribution; genomic location of 21–22 nt reads; genomic location of 25–29 nt reads; heatmap and sequence logo showing T_1 bias; heatmap and sequence logo showing A_{10} bias; barplot showing T_1 bias; 5′ read distance plot for 25–29 nt reads showing enrichment of 10 nt overlap; and a heatmap showing the genomic location of 18–36 bp reads (counts per position: black is low, red is high)

occurs. As a data.frame, this can be easily exported to Excel or other spreadsheet software.

Location-based analyses: Although many viruses are targeted by the siRNA pathway throughout the genome, others are targeted only in limited regions (Sabin *et al.*, 2013). A heatmap representing the occurrence of all mapped read lengths across all genomic locations can be produced using the *size.position.heatmap* function, and barplots showing counts for each genomic location for each read length generated using the *stacked.barplot* function.

Read-based analyses: Read-based analyses allow users to focus on patterns in particular subsets of reads. Single barplots showing the location, strand and count of reads mapping throughout the genome can be visualized using the *position.barplot* function. The base composition of subsets of reads can be calculated with the *make.pwm* function. Sequence signatures of the piRNA pathway include a strong U_1 bias in primary, antisense piRNAs and following 'ping-pong' cycle amplification involving AGO3 and Aub, a strong A_{10} bias in secondary sense piRNAs in *Drosophila* (Brennecke *et al.*, 2007). Similar motifs have been found in piRNAs and viral piRNA-like molecules in mosquitoes or derived cell lines (Morazzani *et al.*, 2012; Schnettler *et al.*, 2013; Vodovar *et al.*, 2012). The output of *make.pwm* can be plotted as a heatmap using the *pwm.heatmap* function, or used with external packages such as seqLogo and motifStack to produce sequence logos. Finally, the 5′-ends of complementary piRNAs are most frequently separated by 10 nt (Brennecke *et al.*, 2007; Vodovar *et al.*, 2012) because of the earlier described 'ping-pong' amplification. The distance between 5′-ends of piRNAs mapping to opposite strands can be summarized and visualized using the *read.dist.plot* function.

3 CONCLUSIONS

Deep sequencing experiments have revealed a variety of interesting and unique signatures of the miRNA, siRNA and piRNA pathways, and there is a need for software that allows scientists to process such data. We have developed viRome, a package for R that allows the interactive generation of a range of informative plots and reports. As an R package, viRome is available on a range of operating systems. viRome is released under an open-source license and can be downloaded from http://virome.sf.net, where a tutorial is also available.

Funding: UK Biotechnology and Biological Sciences Research Council (BBSRC) (BB/J004243/1; BB/J004235/1) (to M.W.); UK Medical Research Council (MRC) (to A.K. and E.S.); The Netherlands Organisation for Scientific Research NWO (Rubicon Fellowship number: 825.10.021) (to E.S.).

Conflict of Interest: none declared.

REFERENCES

Antoniewski,C. (2011) Visitor, an informatic pipeline for analysis of viral siRNA sequencing datasets. *Methods Mol. Biol.*, **721**, 123–142.

Aravin,A.A. *et al.* (2003) The small RNA profile during Drosophila melanogaster development. *Dev. Cell*, **5**, 337–350.

Bernstein,E. *et al.* (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**, 363–366.

Blair,C.D. (2011) Mosquito RNAi is the major innate immune pathway controlling arbovirus infection and transmission. *Future Microbiol.*, **6**, 265–277.

Brennecke,J. *et al.* (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, **128**, 1089–1103.

Donald,C.L. *et al.* (2012) New insights into control of arbovirus replication and spread by insect RNA interference pathways. *Insects*, **3**, 511–531.

Fonseca,N.A. *et al.* (2012) Tools for mapping high-throughput sequencing data. *Bioinformatics*, **28**, 3169–3177.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Lagos-Quintana,M. *et al.* (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Morazzani,E.M. *et al.* (2012) Production of virus-derived ping-pong-dependent piRNA-like small RNAs in the mosquito soma. *PLoS Pathog.*, **8**, e1002470.

Myles,K.M. *et al.* (2009) Origins of alphavirus-derived small RNAs in mosquitoes. *RNA Biol.*, **6**, 387–391.

Sabin,L.R. *et al.* (2013) Dicer-2 processes diverse viral RNA species. *PloS One*, **8**, e55458.

Schnettler,E. *et al.* (2013) RNA interference targets arbovirus replication in culicoides cells. *J. Virol.*, **87**, 2441–2454.

Vodovar,N. *et al.* (2011) In silico reconstruction of viral genomes from small RNAs improves virus-derived small interfering RNA profiling. *J. Virol.*, **85**, 11016–11021.

Vodovar,N. *et al.* (2012) Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. *PloS One*, **7**, e30861.

Zamore,P.D. *et al.* (2000) RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell*, **101**, 25–33.